DATA MINING DAN APLIKASINYA

Contoh Kasus di industri manufaktur dan jasa

Rina Fitriana
Anik Nur Habyba
Elfira Febriani

Sertifikat

Bukti Terbit Buku
No: 256/02/05/xii/2022

Diberikan Kepada :

Dr. Rina Fitriana,ST., MM., IPM.
Anik Nur Habyba, STP., M.Si.
Elfira Febriani, STP., M.Si

Judul Buku :

Data Mining dan Aplikasinya : Contoh Kasus di Industri Manufaktur dan Jasa

ISBN :
978-623-5493-59-6

Yang Diterbitkan di Penerbit Wawasan Ilmu
Pada Tahun 2022

Purwokerto, 28 Desember 2022
Direktur Penerbit Wawasan Ilmu

Nur Wahid

PENERBIT WAWASAN ILMU
No.IKAPI : 215/JTE/2021
Email : naskah.wawasanilmu@gmail.com
WA : 0895 3596 92310
FB : Penerbit Wawasan Ilmu
IG : @penerbitwawasanilmu
Website : www.wawasanilmu.co.id

CV. WAWASAN ILMU

Alamat Kantor: Leler RT 002 RW 006 Desa Kaliwedi Kec. Kebasen Kab. Banyumas Jawa Tengah 53172.
Kantor Layanan : Jl, Martosayogo RT05/02 Teluk, Purwokerto Selatan
No. Hp. +62895359692310, Email: naskah.wawasanilmu@gmail.com, Webs: www.wawasanilmu.co.id
SKT. AHU-0067892-AH.01.14 Tahun 2020
Nomor Anggota IKAPI : 208/JTE/2021

SURAT KETERANGAN
Nomor: 0019/SKT/IX/2022

Yang bertanda tangan di bawah ini menerangkan dengan sesungguhnya bahwa *draft* buku dengan data sebagai berikut:

Judul Buku	: Data Mining dan Aplikasinya : Contoh Kasus di Industri Manufaktur dan Jasa
Tim Penulis	: Dr. Rina Fitriana, ST., MM., IPM., AseanEng

Bahwa draft dengan judul dan atas nama penulis tersebut sedang dalam proses *review* dan terbit:

Nama Penerbit	: CV. Wawasan Ilmu
Kota / Alamat	: Banyumas, Jawa Tengah
Tautan Penerbit	: www.wawasanilmu.co.id
Nomor SKT	: AHU-0067892-AH.01.14 Tahun 2020
Nomor IKAPI	: 215/JTE/2021
Edisi	: Pertama
Bulan/Tahun Terbit	: terjadwal September-Oktober 2022

Demikian Surat Keterangan ini kami buat dengan sebenarnya untuk keperluan penulis sebagaimana mestinya.

Banyumas, 30 September 2022
Hormat kami

Nur Wahid, M.H
Direktur

Data Mining dan Aplikasinya

Contoh Kasus di Industri Manufaktur dan Jasa

Dr.Rina Fitriana,ST., MM., IPM
Anik Nur Habyba, STP., M.Si
Elfira Febriani, STP., M.Si

Data Mining dan Aplikasinya

Contoh Kasus di Industri Manufaktur dan Jasa

WAWASANILMU

Data Mining dan Aplikasinya
Contoh Kasus di Industri Manufaktur dan Jasa

Edisi Pertama
Copyright©2022
WI.2022.0132

Cetakan Pertama: November, 2022
Ukuran: 15,5 cm x 23 cm; Halaman : xxiv + 406

Penulis:
Dr.Rina Fitriana,ST., MM., IPM
Anik Nur Habyba, STP., M.Si
Elfira Febriani, STP., M.Si
Editor : *Nur Wahid*
Cover : Tim Wawasan Ilmu
Tata letak : Nisfi Miftakhul Jannah

Penerbit
Wawasan Ilmu
Anggota IKAPI
Leler RT 002 RW 006 Desa Kaliwedi Kec. Kebasen Kab. Banyumas
Jawa Tengah 53172
Email : naskah.wawasanilmu@gmail.com
Web : https://wawasanilmu.co.id/

ISBN :

PRAKATA

Data Mining merupakan salah satu mata kuliah pilihan di Prodi Teknik Industri di Indonesia. Saat ini masih sedikit sekali buku dalam bahasa Indonesia yang membahas metode-metode Data Mining dengan metode yang berbeda-beda. Pendekatan dan metode-metode untuk Data Mining di bidang Teknik Industri sangat luas, sehingga referensi yang digunakan untuk memenuhi pembelajaran sesuai RPS berasal dari berbagai referensi. Hal ini menyulitkan bagi mahasiswa. Kelebihan buku ajar ini adalah metode-metode yang disampaikan di setiap bab buku ini mewakili beberapa metode yang berbeda-beda sesuai dengan materi dalam RPS yang telah disusun. Buku ini juga dilengkapi dengan contoh soal dan jawaban dan penggunaan software Weka, R, Minitab, Rapid Miner untuk memudahkan pemahaman pembaca terhadap keseluruhan bahasan dalam buku ini. Selain itu juga di setiap bab diberikan penjelasan langkah-langkah penggunaan rumus dan software yang lebih mudah dipahami oleh pembaca.

Mata kuliah pilihan Data Mining pada kurikulum Teknik Industri diberikan untuk mahasiswa S1 dan mk Data Mining dan Aplikasinya adalah mk pilihan pada Program Magister Teknik Industri. Mata kuliah Data Mining juga dikenalkan pada mata kuliah Statistika. Buku ajar ini diharapkan dapat memudahkan dosen di dalam mengajar Data Mining dan memudahkan mahasiswa untuk dapat mempelajari Data Mining. Bab buku ajar ini dibagi sesuai dengan Rencana Pembelajaran Semester (RPS) yang disusun Koordinator mata kuliah Data Mining.

Buku ajar ini diharapkan dapat memudahkan dosen di dalam mengajar Data Mining dan memudahkan mahasiswa untuk dapat mempelajari Data Mining. Data Mining merupakan ilmu yang cukup luas, mencakup banyak metode yang dipelajari khususnya di jurusan Teknik Industri.

Penulis mengucapkan puji dan syukur kepada Allah SWT yang telah memberikan kesempatan kepada penulis untuk dapat menyelesaikan buku ini. Dengan selesainya penulisan buku ajar ini penulis mengucapkan terima kasih kepada :

1. Lembaga Penelitian universitas Trisakti yang telah menyelenggarakan program Hibah Buku Ajar, yang telah mendorong, membantu dan memfasilitasi penulis untuk dapat menyelesaikan buku ini
2. Dosen-dosen di Teknik Industri pada sarjana maupun pasca sarjana yang telah menjadi partner dalam *sharing knowledge* pelaksanaan pengajaran, penelitian dan pengabdian kepada masyarakat dengan topik pengambilan keputusan.
3. Mahasiswa Teknik Industri Universitas Trisakti yang telah membantu menyumbangkan hasil penelitiannya untuk pengayaan buku ini.
4. Pihak lain yang secara langsung dan tidak langsung telah memberikan pengetahuan dan inspirasinya untuk penulisan buku ini.
5. Penerbit yang telah membantu menerbitkan dan memasarkan buku ini.

Buku ini masih jauh dari sempurna, oleh karena itu kami menerima saran dan perbaikan untuk penyempurnaan buku ini.

Jakarta, 30 September 2022

Dr.Rina Fitriana,ST., MM., IPM
Anik Nur Habyba, STP., M.Si
Elfira Febriani, STP., M.Si

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENGANTAR DATA MINING

1.1. Tujuan dan Capaian Pembelajaran

Tujuan:

Mahasiswa mampu menjelaskan dan menguasai konsep data mining, proses pembuatan data mining dan tugas tugas dalam data mining. Bab ini bertujuan untuk mengenalkan prinsip data mining memanfaatkan berbagai sumber belajar, proses pembuatan data mining, jenis jenis tugas data mining dan aplikasinya pada industri manufaktur dan jasa.

Materi yang diberikan :

1. Pengantar Data Mining
2. Proses pembuatan Data Mining
3. Tugas-tugas dalam data mining

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai prinsip Data Mining, mahasiswa mampu menguasai konsep data mining dan aplikasinya. Capaian pembelajaran lulusan adalah mahasiswa mampu menguasai prinsip dan teknik perancangan sistem terintegrasi dengan pendekatan sistem, mampu menguasai teori sains rekayasa, rekayasa perancangan, metode dan teknik terkini yang diperlukan untuk analisis dan perancangan sistem terintegrasi.

1.2. Pengantar Data Mining

Seiring dengan berkembangnya industri 4.0 dan big data, pemerintah mewajibkan adanya mata kuliah yang mendukung industri 4.0 dan big data. Data mining adalah salah satu mata kuliah yang mendukung industri 4.0. Perkembangan dalam teknologi basis data dan tool terotomasi untuk pengumpulan data

telah mengakibatkan menumpuknya data dalam basis data, data warehouses dan tempat penyimpanan data lainnya. Kaya akan data, tapi miskin akan pengetahuan. Solusinya dengan menggunakan analisa data mining yang berupa ekstraksi pengetahuan yang menarik (aturan, pola, atau kendala) dari basis data berukuran besar. Pendekatan dan metode-metode untuk Data Mining dibidang Teknik Industri sangat luas. Hal ini menyulitkan bagi mahasiswa. Kelebihan buku ajar ini adalah metode-metode yang disampaikan di setiap bab buku ini mewakili beberapa metode yang berbeda-beda. Buku ini juga dilengkapi dengan contoh soal dan jawaban dan penggunaan software Weka, R, Minitab, Rapid Miner untuk memudahkan pemahaman pembaca terhadap keseluruhan bahasan dalam buku ini. Selain itu juga di setiap bab diberikan penjelasan langkah-langkah penggunaan rumus dan software yang lebih mudah dipahami oleh pembaca.

Menurut Tan (2006), Data mining adalah proses ekstraksi informasi yang bermanfaat dari data berada pada basis data yang besar yang selama ini tidak diketahui untuk menemukan pola yang berarti dengan cara eksplorasi dan analisis dengan cara otomatis dan semi otomatis. Djatna dan Morimoto (2008) meneliti seleksi atribut basis data numerik berukuran sangat besar berkorelasi dalam lingkungan online. Fattori et al (2003) mengembangkan text mining untuk mengembangkan analisis paten.

Menurut Hand et al (2001) data mining membolehkan menyisir data, memperhatikan pola, merancang aturan, dan membuat prediksi tentang masa depan. Ini dapat didefinisikan sebagai analisa data obervasi (sering yang besar) untuk menentukan hubungan yang tidak terduga dan meringkas data dengan cara baru.

Menurut Gartner Group data *mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2006). *Data mining* memungkinkan pengguna untuk menyaring melalui sejumlah besar informasi yang tersedia di gudang data dari proses pengacakan, dan dilanjutkan dengan BI (Pillai, 2011). Alat data mining yang efisien dan mempresentasikam kerangka sistem BI berdasarkan sparadigma kebanyakan *computational intelligence*, meliputi alat prediksi berdasarkan *neuro-computing* (Jie, 2010).

3

Menurut Moss (2003), *Data Mining* adalah analisis data dengan maksud untuk menemukan permata dari informasi yang tersembunyi dalam jumlah besar data yang telah ditangkap pada saat menjalankan bisnis.

Dalam membuat atau menjalankan *data mining* memerlukan teknik yang gunanya adalah untuk melakukan implementasi spesifik dari algoritma yang digunakan dalam operasi *data mining.* Berikut ini adalah teknik-teknik *data mining* oleh Moss (2003):

1. *Associations Discovery*: teknik *data mining* ini digunakan untuk mengidentifikasi perilaku peristiwa tertentu atau suatu proses. Contoh: Seorang pelanggan *mini market* membeli makanan ringan, ada 85 % kemungkinan bahwa pelanggan tersebut juga akan membeli minuman ringan atau bir.
2. *Sequential pattern discovery*: teknik *data mining* ini mirip dengan asosiasi kecuali pola sekuensial dari waktu ke waktu menentukan bagaimana suatu *item* atau objek berhubungan satu sama lain dari waktu ke waktu. Contoh: Teknik ini dapat memprediksi bahwa seseorang yang membeli mesin cuci juga dapat membeli pengering pakaian dalam waktu enam bulan dengan probabilitas 0,7. Untuk meningkatkan peluang di atas probabilitas 70 % diprediksi, toko mungkin menawarkan setiap pembeli dikson 10 % pada pengering pakaian dalam waktu empat bulan setelah membeli mesin cuci.
3. *Classification*: teknik klasifikasi adalah penggunaan paling umum dari *data mining.* Klasifikasi melihat pada perilaku dan atribut dari kelompok yang telah ditentukan. Contoh: Untuk menemukan karakteristik pelanggan yang akan (atau tidak) membeli jenis produk tertentu, pengetahuan ini akan mengakibatkan berkurangnya biaya promosi dan surat langsung.
4. *Clustering*: teknik ini digunakan untuk menemukan kelompok yang berbeda dalam data. Teknik ini mirip dengan klasifikasi kecuali bahwa ada dari kelompok-kelompok tersebut yang belum ditetapkan pada awal menjalankan *data mining tool*. Contoh: Biasanya teknik ini gunakan untuk mendeteksi masalah seperti cacat manufaktur atau mencari kelompok afinitas untuk kartu kredit.
5. *Forecasting*: Teknik *data mining* ini dibagi menjadi dua bentuk yaitu: Analisis Regresi dan Penemuan Urutan Waktu.

6. Analisis regresi menggunakan nilai-nilai yang diperoleh dari data untuk memprediksi nilai masa depan atau peristiwa masa depan berdasarkan tren historis dan statistik. Misalnya, volume penjualan aksesoris mobil *sport* dapat diperkirakan berdasarkan jumlah mobil *sport* yang dijual bulan lalu.
7. Penemuan urutan waktu berbeda dari analisis regresi dalam hal memperkirakan hanya tergantung pada waktu nilai data. Misalnya, menentukan tingkat kecelakaan selama musim liburan didasarkan pada jumlah kecelakaan yang terjadi selama musim liburan yang sama pada tahun sebelumnya. Properti waktu dapat melingkupi: pekerjaan minggu berdasarkan kalender, liburan, musim, tanggal dan rentang interval tanggal.

Data mining mengaplikasikan algoritma seperti, pohon keputusan , clustering, asosiasi, time series dan sebagainya menjadi suatu data set dan menganalisa isinya. Analisa ini memproduksi pola yang dapat dieksplorasi menjadi informasi yang berguna. Tergantung kepada algoritma yang digarisbawahi, pola ini dapat berupa pohon, rule, aturan , cluster atau formula matematika. Informasi ditemukan pada pola dapat digunakan untuk laporan untuk prediksi. (Tang, 2005)

Data mining dapat diaplikasikan ke dalam tugas berikut ini yaitu (Larose, 2005) klasifikasi, estimasi, segmentasi, asosiasi, peramalan, analisa teks.

Karakteristik utama dari aplikasi data mining adalah :

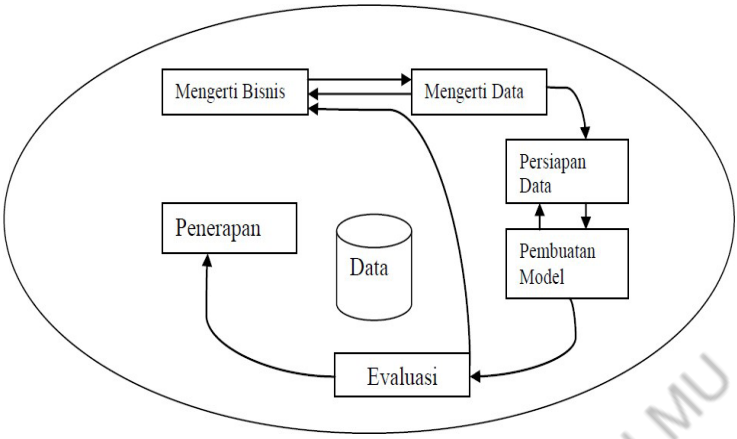
1. Membuat keputusan tanpa coding - Algoritma data mining mempelajari aturan bisnis langsung dari data, membebaskan anda dari percobaan untuk menemukan dan mengkodekan sendiri.
2. Menyesuaikan untuk setiap klien - Data mining mempelajari aturan dari data klien menghasilkan logika yang secara otomatis terspesialisasi untuk setiap individu klien.
3. Otomatis memperbaharuinya - Perubahan bisnis pada klien adalah salah satu faktor yang mempengaruhi bisnis mereka. Data mining membolehkan aplikasi logika secara otomatis terbaharui melalui langkah-langkah proses yang mudah. Aplikasi tidak perlu ditulis ulang, direkompilasi atau diterapkan dan selalu on line.

Teknik data mining dapat diaplikasikan ke banyak aplikasi, menjawab beberapa tipe pertanyaan bisnis, seperti peramalan permintaan, prediksi persediaan, segmentasi produk dan konsumen, manajemen resiko dan lain-lain.

Karakteristik utama dari data mining adalah (Stefanovic, 2009): Membuat keputusan tanpa koding -algoritma data mining mempelajari aturan bisnis langsung dari data, membebaskan Anda untuk menemukan dan mengkodekan sendiri. Diferensiasi dari klien - Data mining mempelajari aturan dari data klien menghasilkan logika yang secara otomatis terspesialisasi untuk masing-masing individu klien. Secara otomatis memperbaharui dirinya sendiri - Perubahan bisnis pada klien adalah faktor yang mempengaruhi bisnisnya. Data mining membolehkan logika aplikasi terbaharui secara terotomatis melalui langkah-langkah proses yang mudah.

Proses pembuatan data mining terdiri dari 6 tahap (Niu, 2009) yaitu:

1. Mengerti Bisnis : Proses pembuatan data mining terdiri mengumpulkan dan mengerti kebutuhan bisnis, meliputi spesifikasi kebutuhan bisnis.
2. Mengerti Data. Data dianalisa pada tahap ini untuk mendeteksi adanya persoalan kualitas data dan bentuk masa depan yang akan dibentuk.
3. Menyiapkan Data. Data di pre-proses untuk mengurangi atau mengeliminasi masalah kualitas data seperti ketidakkonsistenan dan data hilang.
4. Membuat Model. Metode pembuatan spesifikasi data dipilih dan diaplikasikan seperti clustering, spesifikasi, klasifiksi.
5. Evaluasi. Model data dievaluasi dilihat dari sisi perspektif bisnis dan teknis.
6. Penerapan. Jika model data dianggap cukup baik untuk kebutuhan bisnis, maka akan diterapkan dalam lingkungan bisnis untuk membantu pengambilan keputusan.



Gambar 1. 1 Fase Data Mining Model (Niu, 2009)

Data mining sering disebut juga knowledge discovery in database (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Sehingga istilah pattern recognition sekarang jarang digunakan karena ia termasuk bagian dari data mining.

Secara sederhana data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004). Data mining juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007). Data mining, sering juga disebut sebagai knowledge discovery in database (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santoso, 2007). *Data Mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecendrungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti statistic dan matematika (Larose, 2006).

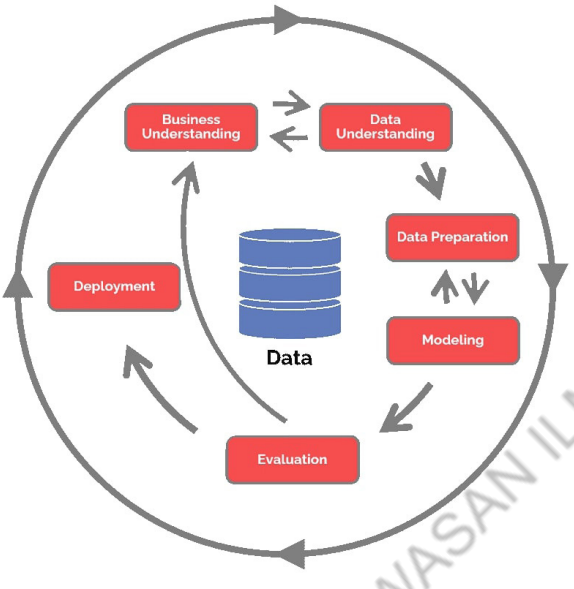
Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, *data warehouse*, atau penyimpanan informasi lainnya. *Data mining*

berkaitan dengan bidang ilmu–ilmu lain, seperti *database system*, *data warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing (Han, 2006). *Data mining* didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar.

Berdasarkan beberapa pengertian tersebut dapat ditarik kesimpulan bahwa *data mining* adalah suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (*database*) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui. Kata *mining* sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligence*), *machine learning*, statistik dan *database*. Beberapa metode yang sering disebut- sebut dalam literatur data mining antara lain *clustering, classification, association rules mining, neural network, genetic algorithm* dan lain-lain (Pramudiono, 2007).

Masalah eksplorasi data adalah perkembangan dalam teknologi basis data dan tool terotomasi untuk pengumpulan data telah mengakibatkan menumpuknya data dalam basis data, *data warehouses* dan tempat penyimpanan data lainnya. Ini menyebabkan perusahaan kaya akan data, tapi miskin akan pengetahuan. Solusi yang diberikan dapat menggunakan Data warehousing dan data mining. Data warehousing dan online analytical processing. Ekstraksi pengetahuan yang menarik (aturan, pola, atau kendala) dari basis data berukuran besar.

CRISP-DM (Cross-Industry Standard Process for Data mining) merupakan sebuah standar di bidang industri, dimana standar ini dibangun sejak tahun 1996 yang digunakan sehingga dapat melakukan proses analisis industri sebagai bentuk strategi pemecahan masalah bisnis (Ayele,2020). CRISP-DM menyediakan standar proses untuk mengekstrak data sebagai strategi pemecahan masalah umum dalam bisnis atau unit penelitian (Martinez-Plumed *et al.*, 2021).



Gambar 1.2 Proses Metode CRISP-DM

Tahapan dari proses data mining dengan menggunakan metode CRISP-DM dapat terlihat pada Gambar, dimana pada siklus tersebut terlihat urutan fase yang tidak kaku dengan adanya pergerakan maju-mundur dan juga tidak searah yang dapat mungkin terjadi pada beberapa fase, hal ini dikarenakan pada hasil dari setiap fase memiliki fungsi tertentu dari fase apa yang harus dilakukan selanjutnya (Damayanti, Kuswayati, 2018). Anak panah menunjukkan dependensi yang paling penting dan sering terjadi di antara fase. Lingkaran luar pada gambar tersebut melambangkan sifat siklus dari data mining itu sendiri, dimana data terus berevolusi dan penambangan data tidak berakhir setelah solusi diterapkan (Bevington, 2003).

1. Business understanding

Pada fase awal ini yaitu pemahaman bisnis, merupakan langkah inti untuk kesuksesan data mining. Pada tahap ini terdapat penganalisisan proyek (Fitriana et.all,2020), pemahaman dan pendefinisian, dengan tujuan agar dapat memahami hal apa yang ingin dicapai pelanggan. Pada pandangan quality management, fase ini diintegrasikan dengan tahapan 'Define', dimana situasi bisnis, proses produksi, dan biasanya permasalahan dapat terjadi, dan tujuan bisnis diidentifikasi, sehingga fase ini berisikan penentuan

objektif dan kebutuhan secara rinci dalam lingkup bisnis. Business understanding juga menilai situasi bisnis, meliputi batasan, asumsi, biaya, resiko, dan keuntungan.

2. Data Understanding

Pada tahapan pemahaman data ini, dilakukan pemahaman mengenai data yang akan diteliti, dengan cara mengumpulkan data. Dalam proses ini data yang ada dapat berasal dari berbagai sumber, sehingga perlu dilakukannya integrasi pada data. Tujuan dari fase ini adalah untuk mengetahui data secara lebih mendalam melalui analisis penyelidikan data untuk mendapatkan informasi awal, mengevaluasi kualitas data dengan memeriksa kevalidan data, dan juga mendeteksipola dari permasalahan melalui data.

3. Data Preparation

Tahapan selanjutnya adalah persiapan data, tahapan ini memiliki tujuan yaitu untuk menyiapkan dataset akhir yang selanjutnya akan dimodelkan. Contoh dari proses yang terdapat pada tahapan persiapan data yaitu menyiapkan kumpulan data yang akan digunakan, dan melakukan seleksi data sesuai dengan analisis yang dilakukan atau data reduction (Bevington, 2003) membersihkan data mentah atau data cleaning (Osborne,2011), serta mengkonstruksikan data menjadi bentuk yang siap untuk dimodelkan atau data transformation.

4. Modelling

Pada tahapan pemodelan, pertama data diproses dengan cara pemilihan teknik pemodelan yang akan digunakan sesuai dengan fungsi data mining, selain itu pemodelan juga dapat digunakan untuk beberapa jenis permasalahan pada data mining yang sama. Pada tahap ini juga memungkinkan untuk kembali pada tahap persiapan data, apabila data dibutuhkan untuk mengolah kembali sehingga dapat menjadi lebih baik sebelum diaplikasikan pada teknik pemodelan.

5. Evaluation

Pada tahap evaluasi, didapatkan hasil dari data mining sesuai kriteria kesuksesan bisnis. Tahap ini dilakukan melalui uji kelayakan serta uji validitas. Selanjutnya juga perlu dilakukan peninjauan ulang terhadap permasalahan bisnis atau penelitian yang tidak ditangani dengan baik, setelah itu mengambil keputusan atau

membuat kebijakan bisnis berkaitan dengan penggunaan hasil dari data mining.

6. Deployment.

Pada tahap ini, mempresentasikan pengetahuan yang diperoleh berdasarkan evaluasi model dari keseluruhan proses data mining, sehingga pada data yang sebelumnya telah dilakukan evaluasi dan modelling, dapat memberikan informasi yang jelas terhadap pengetahuan yang diberikan.

Tugas Data Mining adalah *Classification [Predictive], Clustering [Descriptive], Association Rule Discovery [Descriptive], Sequential Pattern Discovery [Descriptive], Regression [Predictive],Deviation Detection [Predictive].*

Tugas-tugas dalam *Data mining*

Tugas-tugas dalam *data mining* secara umum dibagi ke dalam dua kategori utama:

- Prediktif. Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory* atau variabel bebas.
- Deskriptif. Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend*, *cluster*, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik postprocessing untuk validasi dan penjelasan hasil.

  Berikut adalah tugas-tugas dalam *data mining*:

- Analisis Asosiasi (Korelasi dan kausalitas). Analisis asosiasi adalah pencarian aturan-aturan asosiasi yang menunjukkan kondisi-kondisi nilai atribut yang sering terjadi bersama-sama dalam sekumpulan data. Analisis asosiasi sering digunakan untuk menganalisa *market basket* dan data transaksi. Aturan-aturan asosiasi memiliki bentuk $X \Rightarrow Y$, bahwa $A1 \wedge A2 \wedge \ldots \wedge Am \rightarrow B1 \wedge B2 \wedge \ldots \wedge Bn$, dimana Ai (untuk i = 1, 2, …, m) dan

Bj (untuk j = 1, 2, …, n) adalah pasangan-pasangan nilai atribut. Aturan asosiasi X ⇒ Y diinterpretasikan sebagai tuple-tuple basis data yang memenuhi kondisi kondisi dalam X juga mungkin memenuhi kondisi dalam Y. Contoh dari aturan asosiasi adalah

- ✓ age(X, "20..29") ^ income(X, "20..29K") ⇒ buys(X, "PC") [support = 2%, confidence = 60%]
- ✓ contains(T, "computer") ⇒ contains(x, "software") [1%, 75%]

- Klasifikasi dan Prediksi

  Klasifikasi adalah proses menemukan model (fungsi) yang menjelaskan dan membedakan kelas-kelas atau konsep, dengan tujuan agar model yang diperoleh dapat digunakan untuk memprediksikan kelas atau objek yang memiliki label kelas tidak diketahui. Model yang turunkan didasarkan pada analisis dari training data (yaitu objek data yang memiliki label kelas yang diketahui). Model yang diturunkan dapat direpresentasikan dalam berbagai bentuk seperti aturan IF-THEN klasifikasi, pohon keputusan, formula matematika atau jaringan syaraf tiruan. Dalam banyak kasus, pengguna ingin memprediksikan nilai-nilai data yang tidak tersedia atau hilang (bukan label dari kelas). Dalam kasus ini biasanya nilai data yang akan diprediksi merupakan data numeric. Kasus ini seringkali dirujuk sebagai prediksi. Di samping itu, prediksi lebih menekankan pada identifikasi trend dari distribusi berdasarkan pada data yang tersedia.

- Analisis *Cluster*

  Tidak seperti klasifikasi dan prediksi, yang menganalisis objek data yang diberi label kelas, *clustering* menganalisis objek data dimana label kelas tidak diketahui. *Clustering* dapat digunakan untuk menentukan label kelas tidak diketahui dengan cara mengelompokkan data untuk membentuk kelas baru. Sebab contoh *clustering* rumah untuk menemukan pola distribusinya.

  Prinsip dalam *clustering* adalah memaksimumkan kemiripan *intra-class* dan meminimumkan kemiripan *interclass*.

- Analisis *Outlier*

  Outlier merupakan objek data yang tidak mengikuti perilaku umum dari data. Outlier dapat dianggap sebagai noise atau

pengecualian. Analisis data outlier dinamakan *outlier mining*. Teknik ini berguna dalam *fraud detection* dan *rare events analysis*.

- Analisis Trend dan Evolusi

  Anaisis evolusi data menjelaskan dan memodelkan trend dari objek yang memiliki perilaku yang berubah setiap waktu. Teknik ini dapat meliputi karakterisasi, diskriminasi, asosiasi, klasifikasi, atau *clustering* dari data yang berkaitan dengan waktu. *Data mining* merupakan bidang interdisplin. Disiplin ilmu ini banyak dipengaruhi oleh disiplin sistem basis data, statistika, ilmu informasi, mesin pembelajaran, dan visualisasi. Sistem *data mining* dapat diklasifisikasikan berdasarkan beberapa kategori, yaitu

- Klasifikasi berdasarkan data yang akan di-*mine* seperti *relational*, *transactional*, *object-oriented*, *object-relational*, *spatial*, *time-series*, *text*, multi-media dan www.
- Klasifikasi berdasarkan pengetahuan yang akan di-*mine*, yaitu berdasarkan fungsionalitas *data mining* seperti karakterisasi, diskriminasi, asosiasi, klasifikasi, *clustering*, analisis *outlier* dan analisis evolusi. Sistem *data mining* yang komprehensif biasanya menyediakan beberapa fungsi-fungsi *data mining*.
- Klasifikasi berdasarkan teknik yang akan digunakan seperti *databaseoriented*, *data warehouse* (OLAP), *machine learning*, *Statistics*, *Visualization* dan neural network.
- Klasifikasi berdasarkan aplikasi yang diadaptasi, sebagai contoh system *data mining* untuk keuangan, telekomunikasi, DNA, dan e-mail.

1.3. Latihan Soal

1. Data mining berkaitan dengan penambangan data
   a. Untuk sembarang data set
   b. Data yang berukuran kecil
   c. Data time series saja
   d. Data set berukuran besar
2. Metoda unsupervised diterapkan pada data
   a. Yang sifatnya categorical
   b. Yang sifatnya kontinyus

c. Yang tidak mempunyai label
d. Sembarang data berukuran besar

3. Masuk dalam kelompok unsupervised learning adalah
   a. Kmeans, hirarchical clustering, fuzzy c means
   b. Linear regression, neural networks
   c. Linear discriminant analysis, Support vector machines
   d. K nearest neighbor, Naïve Bayes
4. Jelaskan jenis-jenis data mining dan aplikasinya berdasarkan jurnal dengan Judul A Review of Data Mining Literature pada link berikut ini A_Review_of_Data_Mining_Literature.pdf

BAB II
DATA

2.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Mahasiswa mampu menjelaskan dan menguasai konsep data, tipe tipe data. Bab ini bertujuan untuk mengenalkan pengertian data, tipe tipe data.

Materi yang diberikan :

1. Pengertian data
2. *Non dependency Oriented Data*
3. *Dependency Oriented Data*
4. Data Statistik

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu mengetahui pengertian data, jenis jenis data. Capaian pembelajaran lulusan pada bab ini adalah capaian pembelajaran pengetahuan adalah mampu menguasai prinsip dan teknik perancangan sistem terintegrasi dengan pendekatan sistem.

2.2. Pengertian Data

Data merupakan suatu kumpulan atau catatan fakta-fakta untuk memberikan gambaran yang luas terkait dengan suatu keadaan. Saat ini data dijadikan sebagai suatu acuan untuk memudahkan seseorang mendapatkan informasi setelah data diolah. Melalui data seseorang dapat menganalisis, menggambarkan, atau menjelaskan suatu keadaan. Data juga dibutuhkan di dalam berbagai macam keperluan, seperti penjualan, penelitian hingga kependudukan. Sebagai teknologi umum, data mining dapat diterapkan pada semua jenis data selama data tersebut memiliki makna.

Berdasarkan cara memperolehnya, data dibedakan menjadi dua yaitu **data primer** dan **data sekunder. Data primer** adalah data yang diperoleh serta dikumpulkan secara langsung dari obyek yang diteliti oleh peneliti atau organisasi. Contohnya adalah data hasil survey, data hasil wawancara atau data kuesioner. **Data sekunder** adalah data yang diperoleh dari sumber lain yang telah ada sebelumnya. Peneliti tidak perlu mengumpulkan data secara langsung dari obyek penelitian. Contohnya data sensus penduduk.

Berdasarkan sifat data, data dibedakan menjadi **data kualitatif** dan **data kuantitatif. Data kualitatif** adalah data yang bersifat deskriptif. Data ini dapat diperoleh dari hasil pengisian kuesioner, observasi, wawancara dan lainnya. **Data kuantitatif** adalah data yang diperoleh berupa angka. Contoh data ini adalah umur, tinggi badan dan lainnya.

Salah satu aspek yang menarik dari proses data mining adalah beragamnya tipe data yang tersedia untuk dianalisis. Ada dua jenis data yang luas, dengan kompleksitas yang berbeda-beda, untuk proses data mining:

1. ***Non dependency oriented data***. Tipe data ini biasanya mengacu pada tipe data sederhana seperti data multidimensi atau data teks. Tipe data ini adalah tipe yang paling sederhana dan paling sering ditemui. Dalam kasus ini, data tidak memiliki ketergantungan tertentu antara item data atau atribut. Contohnya adalah kumpulan catatan demografis tentang individu yang berisi usia, jenis kelamin, dan kode pos mereka.
2. ***Dependency oriented data***. Tipe data ini menunjukkan adanya kemungkinan hubungan implisit atau eksplisit yang ada di antara item data. Misalnya, kumpulan data jaringan sosial berisi kumpulan simpul (item data) yang dihubungkan bersama oleh kumpulan sisi (hubungan). Di sisi lain, deret waktu mengandung dependensi implisit. Misalnya, dua nilai berurutan yang dikumpulkan dari sebuah sensor kemungkinan besar terkait satu sama lain.

Secara umum, ***dependency oriented data*** lebih menantang karena kompleksitas yang diciptakan oleh hubungan yang sudah ada sebelumnya antara item data. Ketergantungan seperti itu antara item data perlu dimasukkan langsung ke dalam proses analitis untuk mendapatkan hasil yang bermakna secara kontekstual.

2.3 Non Dependency Oriented Data

Tipe data ini adalah bentuk data yang paling sederhana dan biasanya mengacu pada data multidimensi. Data multidimensi adalah kumpulan n data, X1 . . . Xn, sehingga setiap record $\overline{X}$ i berisi sekumpulan fitur d.

Setiap record berisi sekumpulan field, yang juga disebut sebagai atribut, dimensi, dan fitur. Sistem basis data relasional dirancang secara tradisional untuk menangani jenis data ini, bahkan dalam bentuknya yang paling awal. Sebagai contoh, perhatikan kumpulan data demografi yang diilustrasikan pada Tabel 2.1. Di sini, properti demografis individu, seperti usia, jenis kelamin, dan kode pos, diilustrasikan.

Tabel 2. 1 Contoh Data Set Multidimensional

Nama	Umur	Jenis Kelamin	Suku	Kode Pos
Ahmad	45	Laki-laki	Sunda	10550
Ayu	30	Perempuan	Jawa	10552
Bayu	56	Laki-laki	Jawa	10554
Surya	25	Laki-laki	Batak	13790

2.3.1. Quantitative Multidimensional Data

Atribut pada Tabel 1.1 terdiri dari dua jenis yang berbeda. Bidang usia memiliki nilai numerik dan dapat disebut atribut tersebut sebagai data yang bersifat kontinu, numerik, atau kuantitatif. Data yang semua bidangnya bersifat kuantitatif disebut juga sebagai data kuantitatif atau data numerik. Setiap nilai yang bersifat kuantitatif, maka kumpulan data yang sesuai disebut sebagai data multidimensi kuantitatif.

Subtipe ini sangat cocok untuk pemrosesan analitik karena jauh lebih mudah untuk mengolah data kuantitatif dari perspektif statistik. Misalnya, rata-rata dari satu set data kuantitatif dapat dinyatakan sebagai rata-rata sederhana dari nilai-nilai ini, sedangkan perhitungan tersebut menjadi lebih kompleks dalam tipe data lainnya. Jika memungkinkan dan efektif, banyak algoritme data mining mencoba mengubah berbagai jenis data menjadi nilai kuantitatif sebelum diproses. Namun demikian, dalam aplikasi nyata, data cenderung lebih kompleks dan mungkin berisi campuran dari tipe data yang berbeda.

2.3.2. Data Atribut Kategoris

Banyak kumpulan data dalam aplikasi nyata mungkin berisi atribut kategoris yang mengambil nilai diskrit tidak berurutan. Misalnya, pada Tabel 1.1, atribut seperti jenis kelamin, ras, dan kode pos maka data tersebut disebut sebagai bernilai diskrit tak beraturan atau kategorikal. Dalam kasus data atribut campuran, ada kombinasi atribut kategorikal dan numerik. Data lengkap pada Tabel 1.1 dianggap data atribut campuran karena mengandung atribut numerik dan kategorik.

Atribut pada nilai gender bersifat kategoris, tetapi hanya dengan dua kemungkinan nilai. Atribut ini dapat disebut sebagai data biner, dan dapat dianggap sebagai kasus khusus dari data numerik atau kategorikal.

2.3.3. Binary Data

Angka biner sering digunakan ketika mengoperasikan computer. Bilangan biner adalah jenis penulisan angka menggunakan symbol yaitu 0 dan 1. Bilangan biner adalah dasar dari semua bilangan berbasis digital. Pengelompokan biner dalam istilah computer selalu berjumlah 8 dengan istilah 1 byte.

Data biner dapat dianggap sebagai kasus khusus dari data kategorikal multidimensi atau data kuantitatif multidimensi. Ini adalah kasus khusus dari data kategorikal multidimensi, di mana setiap atribut kategorikal dapat mengambil salah satu dari paling banyak dua nilai diskrit. Data biner juga merupakan representasi dari data setwise, di mana setiap atribut diperlakukan sebagai indikator elemen set. Nilai 1 menunjukkan bahwa elemen harus dimasukkan dalam himpunan.

2.3.4. Data Teks

Data teks adalah salah satu jenis data yang muncul setelah lahirnya era big data. Data teks dapat dilihat baik sebagai string, atau sebagai data multidimensi. Setiap string adalah urutan karakter (atau kata) yang sesuai dengan dokumen.

Salah satu contoh penerapan penggunaan data teks, misalnya review dari suatu produk untuk menentukan apakah produk tersebut memiliki review positif atau malah sebaliknya. Selain itu, tweet yang dilakukan oleh jutaan orang mengenai topik tertentu juga dapat dijadikan contoh. Proses pengolahan data yang

digunakan untuk menganalisis data teks tidak akan sama dengan cara mengolah data yang berbentuk angka.

2.4. Dependency Oriented Data

Dalam praktiknya, nilai data yang berbeda mungkin (secara implisit) terkait satu sama lain secara temporal, spasial, atau melalui tautan hubungan jaringan eksplisit antara item data. Perkembangan pengetahuan tentang ketergantungan data yang sudah ada sebelumnya sangat mengubah proses analisis data mining karena data mining adalah tentang menemukan hubungan antara item data. Kehadiran dependensi data mengubah hubungan yang diharapkan dalam data, dan apa yang mungkin dianggap menarik dari perspektif hubungan data tersebut. Beberapa jenis dependensi data adalah sebagai berikut:

a. Ketergantungan implisit: Dalam hal ini, dependensi antara item data tidak ditentukan secara eksplisit tetapi diketahui "biasanya" ada di domain itu. Misalnya, nilai suhu berurutan yang dikumpulkan oleh sensor cenderung sangat mirip satu sama lain. Oleh karena itu, jika nilai suhu yang direkam oleh sensor pada waktu tertentu berbeda secara signifikan dari yang direkam pada waktu b. berikutnya, maka ini sangat tidak biasa dan mungkin menarik untuk proses data mining.
b. Ketergantungan eksplisit: Ini biasanya mengacu pada grafik atau data jaringan. Grafik adalah abstraksi yang sangat kuat yang sering digunakan sebagai representasi perantara untuk memecahkan masalah data mining dalam konteks tipe data lainnya.

2.4.1. Data Time Series

Time series atau runtun waktu adalah himpunan observasi data dalam waktu. Data time series atau data runtun waktu berisi nilai yang biasanya dihasilkan oleh pengukuran berkelanjutan dari waktu ke waktu. Misalnya, sensor lingkungan akan mengukur suhu secara terus-menerus, sedangkan elektrokardiogram (EKG) akan mengukur parameter irama jantung seseorang. Data tersebut biasanya memiliki ketergantungan implisit pada nilai yang diterima dari waktu ke waktu. Misalnya, nilai yang berdekatan yang direkam oleh sensor suhu biasanya akan bervariasi dari waktu ke waktu, dan faktor ini perlu digunakan secara eksplisit dalam proses data mining.

Tidak seperti data multidimensi dimana semua atribut diperlakukan sama, data time series merupakan representasi dari data kontekstual. Oleh sebab itu, atribut diklasifikasikan menjadi dua jenis:

1. Atribut kontekstual: Ini adalah atribut yang mendefinisikan konteks dimana pemukuran dilakukan. Atribut ini memberikan pengetahuan dimana nilai-nilai perilaku diukur. Beberapa tipe data seperti data spasial berisi atribut kontekstual yang sesuai dengan koordinat spasial. Misalnya, dalam kasus data sensor, cap atau stemple waktu di mana pembacaan diukur dapat dianggap sebagai atribut kontekstual.
2. Atribut perilaku: Ini mewakili nilai-nilai yang diukur dalam konteks tertentu. Dalam contoh sensor, suhu adalah nilai atribut perilaku. Atribut ini juga dimungkinkan untuk memiliki lebih dari satu atribut perilaku. Misalnya, jika beberapa sensor merekam pembacaan pada stempel waktu yang disinkronkan, maka itu menghasilkan kumpulan data deret waktu multidimensi.

2.4.2. Discrete Sequences and Strings

Discrete sequences atau urutan diskrit dapat dianggap sebagai analog kategoris dari data deret waktu. Seperti dalam kasus data time series, atribut kontekstual adalah cap waktu atau indeks posisi dalam pengurutan. Atribut perilaku adalah nilai kategoris. Oleh karena itu, data urutan diskrit didefinisikan dengan cara yang mirip dengan data time series.

Misalnya, dalam kasus akses Web, di mana alamat halaman Web dan alamat IP asal permintaan dikumpulkan untuk 100 akses berbeda. Ini mewakili urutan diskrit dengan panjang n = 100 dan dimensi d = 2. Kasus yang sangat umum dalam data sequences adalah skenario univariat, di mana nilai d adalah 1. Data urutan seperti itu juga disebut sebagai string.

Perlu dicatat bahwa definisi tersebut di atas hampir identik dengan kasus time series, dengan perbedaan utama adalah bahwa urutan diskrit mengandung atribut kategoris. Secara teori, sangat memungkinkan untuk memiliki data antara data kategorikal dan numerik. Contoh lainnya adalah transaksi supermarket mungkin berisi urutan set item. Setiap set dapat berisi sejumlah item. Urutan setwise seperti ini sebenarnya bukan urutan multivariat, tetapi urutan univariat, di mana setiap elemen dari urutan adalah

himpunan yang bertentangan dengan elemen unit. Oleh sebab itu discrete sequences dapat didefinisikan dalam berbagai cara yang lebih luas, dibandingkan dengan data time series karena kemampuan untuk mendefinisikan himpunan pada elemen diskrit.

Dalam beberapa kasus, atribut kontekstual mungkin tidak merujuk ke waktu secara eksplisit, tetapi posisi berdasarkan penempatan fisik. Kasus seperti ini dapat ditemukan dalam sekuens biologis. Beberapa contoh skenario umum di mana data sequences mungkin muncul adalah sebagai berikut:

1. Log peristiwa: Berbagai macam sistem komputer, server Web, dan aplikasi Web membuat log peristiwa berdasarkan aktivitas pengguna. Contoh log peristiwa adalah urutan tindakan pengguna di situs Web keuangan:

   Login Password Login Password Login Password…

   Urutan khusus ini dapat mewakili skenario di mana pengguna mencoba untuk membobol sistem yang dilindungi kata sandi

2. Data biologis: Dalam hal ini, urutannya mungkin sesuai dengan rangkaian nukleotida atau asam amino. Urutan unit tersebut memberikan informasi tentang karakteristik fungsi protein. Oleh karena itu, proses data mining dapat digunakan untuk menentukan pola menarik yang mencerminkan sifat biologis yang berbeda.

2.4.3. Data Spasial

Data spasial adalah data yang menunjukkan lokasi letak data tersebut. Data spasial memiliki referensi posisi geografis dan digambarkan dalam sebuah sistem koordinat. Seiring berkembangnya produksi data, maka jumlah data spasial juga bertambah dengan pesat.

Dalam data spasial, banyak atribut nonspasial (misalnya, suhu, tekanan, intensitas warna piksel gambar) diukur di lokasi spasial. Misalnya, suhu permukaan laut sering dikumpulkan oleh ahli meteorologi untuk meramalkan terjadinya angin topan. Dalam kasus seperti itu, koordinat spasial sesuai dengan atribut kontekstual, sedangkan atribut seperti suhu sesuai dengan atribut perilaku. Biasanya, ada dua atribut spasial. Seperti dalam kasus data time series, juga dimungkinkan untuk memiliki beberapa atribut perilaku. Misalnya, dalam aplikasi suhu permukaan laut, seseorang mungkin juga mengukur atribut perilaku lainnya seperti tekanan.

Definisi tersebut di atas memberikan fleksibilitas yang luas dalam hal bagaimana record Xi dan lokasi Li dapat didefinisikan. Misalnya, atribut perilaku dalam catatan Xi mungkin numerik atau kategoris, atau campuran keduanya. Dalam aplikasi meteorologi, Xi mungkin berisi atribut suhu dan tekanan di lokasi Li. Selanjutnya, Li dapat ditentukan dalam hal koordinat spasial yang tepat, seperti lintang dan bujur, atau dalam hal lokasi logis, seperti kota atau negara bagian.

Data mining untuk data spasial terkait erat dengan penambangan data time series, di mana atribut perilaku dalam aplikasi spasial yang paling umum dipelajari adalah kontinu, meskipun beberapa aplikasi dapat menggunakan atribut kategoris juga. Oleh karena itu, kontinuitas nilai diamati di seluruh lokasi spasial yang berdekatan, seperti halnya kontinuitas nilai yang diamati di seluruh stempel waktu yang berdekatan dalam data time series.

2.4.4. Data Jaringan dan Grafik

Dalam jaringan dan data grafik, nilai data mungkin sesuai dengan node dalam jaringan, sedangkan hubungan antara nilai data mungkin sesuai dengan tepi dalam jaringan. Dalam beberapa kasus, atribut dapat dikaitkan dengan node dalam jaringan. Meskipun dimungkinkan juga untuk mengasosiasikan atribut dengan edge dalam jaringan, hal ini jauh lebih jarang dilakukan.

Grafik berada dimana-mana dalam berbagai domain aplikasi seperti kimia, data biologis, bioinformatika dan bidang lainnya. Banyak informasi penting dari grafik yang dapat diperoleh untuk dimanfaatkan dan dianalisis. Dalam aplikasi seperti data kimia dan biologi, database tersedia dari banyak grafik-grafik di mana setiap node dikaitkan dengan label. Misalnya, grafik Web mungkin berisi node terarah yang sesuai dengan arah hyperlink antar halaman, sedangkan pertemanan di jejaring sosial Facebook tidak terarah. Pada jejaring sosial, simpul berhubungan dengan aktor jaringan sosial, sedangkan ujungnya berhubungan dengan hubungan pertemanan. Node mungkin memiliki atribut yang sesuai dengan konten halaman sosial. Dalam beberapa bentuk khusus dari jaringan sosial, seperti email atau jaringan chat-messenger, edge mungkin memiliki konten yang terkait dengannya. Konten ini sesuai dengan komunikasi antara node yang berbeda.

Data jaringan atau *network* adalah representasi yang sangat umum dan dapat digunakan untuk menyelesaikan banyak

aplikasi berbasis kesamaan pada tipe data lain. Sebagai contoh, data multidimensi dapat diubah menjadi data jaringan dengan membuat sebuah node untuk setiap record dalam database, dan merepresentasikan kesamaan antara node dengan edge. Representasi seperti itu cukup sering digunakan untuk banyak aplikasi data mining berbasis kesamaan, seperti pengelompokan. Hal ini dimungkinkan untuk menggunakan algoritma deteksi komunitas untuk menentukan cluster dalam data jaringan dan kemudian memetakannya kembali ke data multidimensi.

2.5. Data Statistik

Sebelum melakukan analisis maka penting untuk memiliki gambaran keseluruhan tentang data. Penggunaan statistik dasar dapat digunakan untuk mengidentifikasi properti data dan menyoroti nilai data mana yang harus diperlakukan sebagai noise atau outlier.

Bagian ini membahas dua bidang deskripsi statistik dasar. Bagian ini dimulai dengan ukuran *measures of central tendency* yang mengukur pusat distribusi data dimana pembahasan meliputi mean, median, modus, dan midrange.

Selain *measures of central tendency*, terdapat juga pembahasa mengenai penyebaran data. Artinya, bagaimana data itu tersebar? Ukuran dispersi data yang paling umum adalah jangkauan, kuartil, dan jangkauan interkuartil; boxplot; dan varians dan standar deviasi data. Langkah-langkah ini berguna untuk mengidentifikasi outlier.

2.5.1. Tendensi Sentral: Mean, Median, dan Modus

Tendensi sentral merupakan ukuran statistic untuk menentukan skor tunggal yang dapat digunakan untuk mendefinisikan pusat distribusi. Jenisnya adalah mean (rata-rata), median (nilai tengah), dan Modus. **Mean** merupakan angka rata-rata atau perhitungan rata-rata dari suatu data. Misalkan x1,x2,...,xN adalah himpunan nilai atau observasi N, seperti untuk beberapa atribut numerik X, seperti tinggi tanaman. Perhitungan mean dari himpunan nilai ini adalah

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Contoh :

Misalkan terdapat suatu data tinggi tanaman penelitian berikut (dalam cm), ditunjukkan dalam urutan yang meningkat: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Menggunakan Persamaan. (2.1), maka:

$$\overline{X} = \frac{30 + 36 + 47 + 50 + 52 + \cdots + 110}{12} = 58$$

Sehingga rata-rata dari tinggi tanaman tersebut adalah 58 cm.

Kadang-kadang, setiap nilai xi dalam suatu himpunan dapat dikaitkan dengan bobot wi untuk i = 1,...,N. Bobot mencerminkan signifikansi, kepentingan, atau frekuensi kejadian yang melekat pada nilai masing-masing, disebut dengan ***weighted average***. Dalam hal ini, persamaan ini menjadi:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}$$

Meskipun mean adalah kuantitas tunggal yang paling berguna untuk menggambarkan kumpulan data, namun mean tidak selalu merupakan cara terbaik untuk mengukur pusat data. Masalah utama dengan mean adalah kepekaannya terhadap nilai ekstrim (misalnya, outlier). Bahkan sejumlah kecil nilai ekstrim dapat merusak mean. Misalnya, gaji rata-rata disebuah perusahaan mungkin secara substansial didorong oleh beberapa manajer yang dibayar tinggi. Demikian pula, nilai rata-rata suatu kelas dalam ujian dapat diturunkan sedikit oleh beberapa nilai yang sangat rendah. Untuk mengimbangi efek yang disebabkan oleh sejumlah kecil nilai ekstrim, maka dapat menggunakan trimmed mean, yaitu mean yang diperoleh setelah memotong nilai pada ekstrim tinggi dan rendah.

Misalnya, hal yang dapat dilakukan dengan mengurutkan nilai yang diamati untuk tinggi tanaman dan menghapus 2% atas dan bawah sebelum menghitung rata-rata. Namun tetap harus diperhatikan untuk tidak melakukan pemangkasan porsi yang terlalu besar (seperti 20%) di kedua ujungnya, karena hal ini dapat mengakibatkan hilangnya informasi yang berharga.

Untuk data miring (asimetris), ukuran pusat data yang lebih baik adalah **median**, yang merupakan nilai tengah dalam kumpulan

nilai data terurut. Ini adalah nilai yang memisahkan bagian yang lebih tinggi dari kumpulan data dari bagian yang lebih rendah.

Dalam probabilitas dan statistik, median umumnya berlaku untuk data numerik; namun, dapat diperluas ke konsep data ordinal. Misalkan kumpulan data yang diberikan nilai N untuk atribut X diurutkan dalam urutan yang meningkat. Jika N ganjil, maka median adalah nilai tengah dari himpunan terurut. Jika N genap, maka median tidak tunggal; itu adalah dua nilai paling tengah dan nilai apa pun di antaranya. Jika X adalah atribut numerik dalam hal ini, menurut konvensi, median diambil sebagai rata-rata dari dua nilai paling tengah.

Contoh:

Berdasarkan contoh sebelumnya, data sudah diurutkan dalam urutan nilai dari yang paling kecil hingga paling besar. Jumlah data pengamatan genap (yaitu, 12). Karena berjumlah genap maka nilai yang diambil adalah pada data ke 6 dan data ke 7 (nilai 52 dan 56). Sehingga nilai sebagai median; yaitu, 52 + 56 = 108/2 = 54. Jadi, mediannya adalah 54 cm.

Misalkan data pada contoh tersebut bernilai ganjil (data 1 hingga data 11) maka median adalah nilai paling tengah. Jika nilai paling tengah berarti median pada data keenam berupa 52 cm.

Modus adalah ukuran lain dari tendensi sentral. Modus untuk sekumpulan data adalah nilai yang paling sering muncul dalam himpunan tersebut. Oleh karena itu, dapat ditentukan atribut kualitatif dan kuantitatif. Frekuensi terbesar dimungkinkan untuk berhubungan dengan beberapa nilai yang berbeda, yang menghasilkan lebih dari satu mode. Kumpulan data dengan satu, dua, atau tiga mode masing-masing disebut unimodal, bimodal, dan trimodal. Secara umum, kumpulan data dengan dua atau lebih mode adalah multimodal. Di sisi lain, jika setiap nilai data hanya muncul sekali, maka tidak ada mode.

Contoh:

Berdasarkan data sebelumnya maka modus atau nilai yang paling sering muncul adalah 52 cm dan 70 cm.

Midrange juga dapat digunakan untuk menilai tendensi sentral dari kumpulan data numerik. Ini adalah rata-rata dari nilai terbesar dan terkecil dalam himpunan. Ukuran ini mudah dihitung menggunakan fungsi agregat SQL, max () dan min().

Contoh:

Midrange berdasarkan contoh sebelumnya adalah nilai minimum (30) + nilai maksimum (110) = 140/2 = 70 cm.

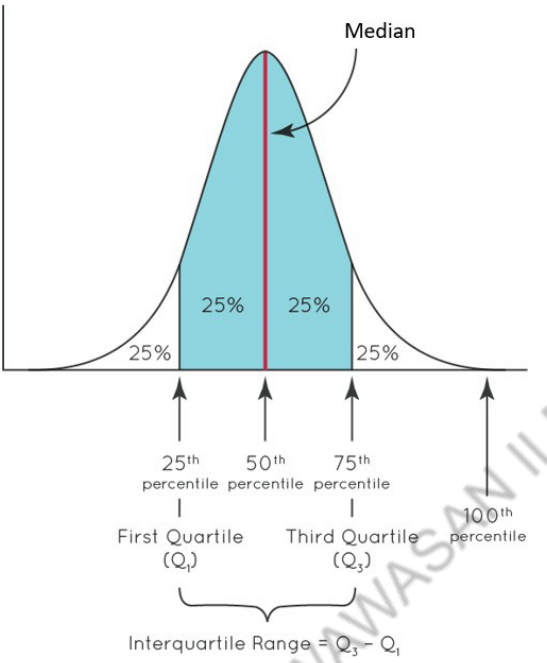
2.5.2. Data Dispersi: Range, Quartile, Variance, dan Standar Deviasi

Misalkan x1,x2,...,xN adalah himpunan pengamatan untuk beberapa atribut numerik, X. Jangkauan himpunan adalah selisih antara nilai terbesar (maks()) dan terkecil (min()). Misalkan data untuk atribut X diurutkan dalam urutan numerik dari nilai terkecil hingga nilai tertinggi. Kemudian data tersebut dibagi distribusi data menjadi kumpulan-kumpulan yang berurutan dengan ukuran yang sama. Titik data ini disebut **kuantil**.

Kuantil adalah titik yang diambil pada interval dari distribusi data, membaginya menjadi set berurutan yang jumlahnya sama. Kuartil q ke-k untuk distribusi data yang diberikan adalah nilai x sedemikian rupa sehingga paling banyak k/q dari nilai data lebih kecil dari x dan paling banyak (q k)/q dari nilai data lebih dari x, di mana k adalah bilangan bulat sehingga 0 < k < q. Ada q 1 q-kuantil.

Dua kuantil adalah titik data yang membagi bagian bawah dan atas dari distribusi data. Hal tersebut sama seperti median. Empat kuartil adalah tiga titik data yang membagi distribusi data menjadi empat bagian yang sama; setiap bagian mewakili seperempat dari distribusi data. Hal ini lebih sering disebut sebagai **kuartil**. 100-kuantil lebih sering disebut sebagai **persentil**; mereka membagi distribusi data menjadi 100 set berurutan berukuran sama. Median, kuartil, dan persentil adalah bentuk kuantil yang paling banyak digunakan.

Kuartil memberikan indikasi pusat distribusi, penyebaran, dan bentuk. Kuartil pertama, dilambangkan dengan Q1, adalah persentil ke-25. Ini memotong 25% terendah dari data. Kuartil ketiga, dilambangkan dengan Q3, adalah persentil ke-75—memotong 75% terendah (atau 25% tertinggi) dari data. Kuartil kedua adalah persentil ke-50. Sebagai median, itu memberikan pusat distribusi data. Pembagian ini dapat dilihat pada Gambar 2.1.



Gambar 2. 1 Grafik untuk Kuartil Q1, Q2, dan Q3

Jarak antara kuartil pertama dan ketiga adalah ukuran penyebaran sederhana yang memberikan rentang yang dicakup oleh bagian tengah data. Jarak ini disebut jangkauan interkuartil (IQR) dan didefinisikan sebagai:

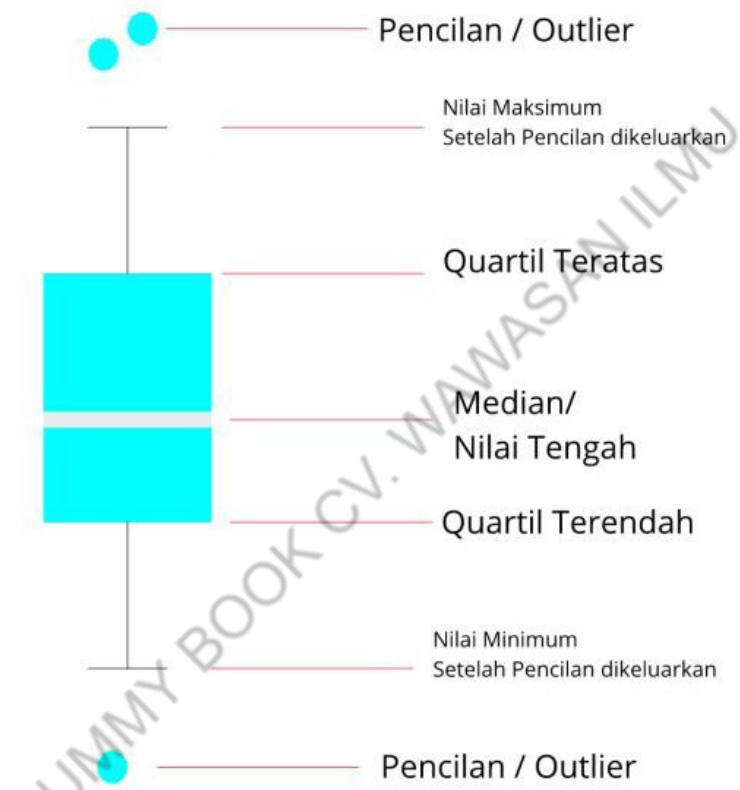
$$IQR = Q_3 - Q_1$$

Contoh

Kuartil adalah tiga nilai yang membagi kumpulan data yang diurutkan menjadi empat bagian yang sama. Berdasarkan contoh sebelumnya terdapat 12 pengamatan yang sudah diurutkan dalam urutan nilai dari yang terkecil hingga terbesar. Jadi, kuartil untuk data ini masing-masing adalah nilai ketiga, keenam, dan kesembilan, dalam daftar yang diurutkan. Oleh karena itu, Q1 = 47 cm dan Q3 adalah 63 cm. Jadi, jangkauan interkuartilnya adalah IQR = 63-47 = 16 cm. (Perhatikan bahwa nilai keenam adalah median, 52 cm, meskipun kumpulan data ini memiliki dua median karena jumlah nilai datanya genap).

2.5.3. Boxplot dan Outlier

Boxplot adalah cara populer untuk memvisualisasikan distribusi. Boxplot merupakan ringkasan distribusi sampel yang disajikan secara grafis yang dapat menggambarkan bentuk distribusi data (*skewness*). Sebuah boxplot menggabungkan beberapa hal sebagai berikut (Gambar 2.2):



Gambar 2. 2 Penjelasan Boxplot

1. Bagian utama boxplot adalah kotak berbentuk persegi (**Box**) yang merupakan bidang yang menyajikan interquartile range (**IQR**), dimana 50 % dari nilai data pengamatan terletak di sana.
   - Panjang kotak (box) sesuai dengan jangkauan kuartil dalam (IQR). Semakin panjang bidang IQR menunjukkan data semakin menyebar.
   - Garis bawah kotak (LQ) = Q1, dimana 25% data pengamatan lebih kecil atau sama dengan nilai Q1.

- Garis tengah kotak = Q2 (median), dimana 50% data pengamatan lebih kecil atau sama dengan nilai ini.
- Garis atas kotak (**UQ**) = Q3 (Kuartil ketiga) dimana 75% data pengamatan lebih kecil atau sama dengan nilai Q1.

2. Garis yang merupakan perpanjangan dari **box** (baik ke arah atas ataupun ke arah bawah) disebut sebagai **whiskers**.
   - Whiskers bawah menunjukkan nilai yang lebih rendah dari kumpulan data yang berada dalam IQR.
   - Whiskers atas menunjukkan nilai yang lebih tinggi dari kumpulan data yang berada dalam IQR.
   - Panjang whisker ≤ 1.5 x IQR. Masing-masing garis whisker dimulai dari ujung kotak IQR, dan berakhir pada nilai data yang bukan dikategorikan sebagai outlier (*Pada gambar, batasnya adalah garis UIF dan LIF*).
3. **Outlier** adalah nilai data yang letaknya melebihi IQR, diukur dari UQ (atas kotak) atau LQ (bawah kotak). Nilai outlier dapat dilihat dari:
   - Q3 + (1.5 x IQR) < outlier atas ≤ Q3 + (3 x IQR)
   - Q1 – (1.5 x IQR) > outlier bawah ≥ Q1 – (3 x IQR)
4. **Nilai ekstrim** adalah nilai-nilai yang letaknya lebih dari 3 x panjang kotak (IQR), diukur dari UQ (atas kotak) atau LQ (bawah kotak).
   - Ekstrim bagian atas apabila nilainya berada di atas Q3 + (3 x IQR) dan
   - Ekstrim bagian bawah apabila nilainya lebih rendah dari Q1 – (3 x IQR)

2.5.4. Varians dan Standar Deviasi

Varians dan standar deviasi adalah ukuran dispersi data. Varians dan standar deviasi menunjukkan seberapa menyebar distribusi data. Standar deviasi yang rendah berarti bahwa pengamatan data cenderung sangat dekat dengan mean, sedangkan standar deviasi yang tinggi menunjukkan bahwa data tersebar pada rentang nilai yang besar.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Dimana $\overline{X}$ adalah rata-rata pengamatan, Standar deviasi, σ, adalah akar kuadrat dari varians, σ^2

Contoh:

Pada contoh sebelumnya diketahui bahwa rata-rata (mean) pengamatan sebesar 58 cm. Untuk menentukan varians dan standar deviasi data dari contoh itu, dengan N = 12 didapatkan:

$$\sigma^2 = \frac{1}{12} = (30^2 + 36^2 + \cdots + 110^2) - 58^2 = 379{,}17$$

Kemudian

$\sigma = \sqrt{379{,}17} = 19{,}47$

Sifat dasar standar deviasi sebagai ukuran penyebaran adalah sebagai berikut:

5. σ mengukur sebaran tentang mean dan harus dipertimbangkan hanya jika mean dipilih sebagai ukuran pusat.
6. $\sigma = 0$ hanya jika tidak ada sebaran, yaitu jika semua pengamatan memiliki nilai yang sama. Jika tidak, $\sigma > 0$.

2.5.5. Quantile-Quantile Plot

Quantile-quantile plot, atau **q-q plot**, merupakan grafik kuantil dari satu distribusi univariat terhadap kuantil yang sesuai dari yang lain. Q-Q plot adalah alat visualisasi yang kuat karena memungkinkan pengguna untuk melihat apakah ada pergeseran dari satu distribusi ke distribusi lainnya.

Misalkan kita memiliki dua set pengamatan untuk atribut atau harga satuan variabel, yang diambil dari dua lokasi cabang yang berbeda. Biarkan $x_1 , \ldots , x_N$ adalah data dari cabang pertama, dan $y_1,\ldots,y_M$ adalah data dari cabang kedua, di mana setiap kumpulan data diurutkan dalam urutan yang meningkat. Jika M = N (yaitu, jumlah titik di setiap himpunan adalah sama), maka cukup memplot yi terhadap xi , di mana yi dan xi keduanya (i - 0,5)/N kuantil dari kumpulan data masing-masing. Jika M < N (yaitu, cabang kedua memiliki pengamatan lebih sedikit daripada yang pertama), hanya ada titik M pada plot q-q. Di sini, yi adalah kuantil (i - 0,5)/M dari data y, yang diplot terhadap kuantil (i - 0,5)/M dari data x. Perhitungan ini biasanya melibatkan interpolasi.

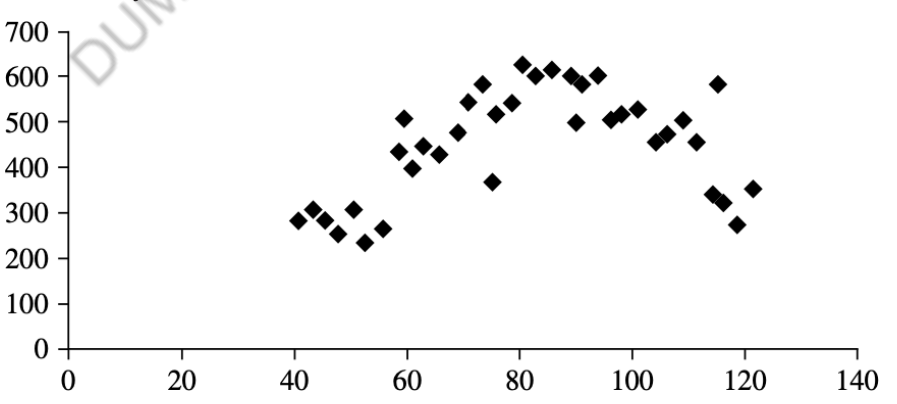
2.5.6. Histogram

Histogram telah banyak digunakan dalam pengolahan data. "Histos" berarti tiang, dan "gram" berarti bagan, jadi **histogram** adalah tampilan berbentuk grafis untuk menunjukkan distribusi data secara visual atau seberapa sering suatu nilai yang berbeda terjadi pada suatu kumpulan data. Jika X adalah nominal, seperti model mobil atau jenis barang, maka tiang atau batang vertikal digambar untuk setiap nilai X yang diketahui. Tinggi batang menunjukkan frekuensi (yaitu, hitungan) dari nilai X itu. Grafik yang dihasilkan lebih dikenal sebagai diagram batang.

Jika X numerik, istilah histogram lebih disukai. Rentang nilai untuk X dipartisi ke dalam subrentang berurutan yang terputus-putus. Subrange, disebut sebagai bucket atau bin, adalah subset terpisah dari distribusi data untuk X. Rentang bucket dikenal sebagai lebar. Misalnya, atribut suatu nilai dengan rentang nilai 1 hingga 200 dapat dipartisi menjadi subrentang 1 hingga 20, 21 hingga 40, 41 hingga 60, dan seterusnya. Untuk setiap subrange, sebuah bar digambar dengan ketinggian yang mewakili jumlah total item yang diamati dalam subrange.

2.5.7. Scatter Plot dan Korelasi Data

Scatter plot adalah salah satu metode grafis yang paling efektif untuk menentukan apakah terdapat hubungan, pola, atau tren antara dua atribut numerik. Untuk membangun scatter plot, setiap pasangan nilai diperlakukan sebagai pasangan koordinat dalam pengertian aljabar dan diplot sebagai titik pada bidang. Pada Gambar 2.3. menunjukkan grafik scatter plot berdasarkan contoh sebelumnya.



Gambar 2. 3 Scatter plot

Scatter plot adalah metode yang berguna untuk memberikan pandangan pertama pada data bivariat untuk melihat kelompok titik dan outlier, atau untuk mengeksplorasi kemungkinan hubungan korelasi. Dua atribut, X, dan Y, berkorelasi jika satu atribut menyiratkan yang lain. Korelasi bisa positif, negatif, atau nol (tidak berkorelasi). Gambar 2.4 menunjukkan contoh korelasi positif dan negatif antara dua atribut.

Gambar 2. 4 Scatter plot untuk korelasi positif (a) dan korelasi negative (b)

Jika pola titik yang diplot miring dari kiri bawah ke kanan atas, ini berarti bahwa nilai X meningkat seiring dengan peningkatan nilai Y, menunjukkan korelasi positif (Gambar 2.4a). Jika pola titik-titik yang diplot miring dari kiri atas ke kanan bawah, nilai X meningkat ketika nilai Y menurun, menunjukkan korelasi negatif (Gambar 2.4b). Garis yang paling cocok dapat ditarik untuk mempelajari korelasi antar variabel.

BAB III
EKSPLORASI DATA

3.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Mahasiswa mampu menjelaskan dan menguasai pengertian eksplorasi data. Bab ini bertujuan untuk mengenalkan konsep eksplorasi data

Materi yang diberikan :

1. Pengertian Eksplorasi data
2. Pembersihan data
3. *Noisy Data*
4. Nilai yang hilang
5. Data Integrasi
6. Transformasi Data
7. Reduksi Data
8. Analisis Korelasi dan Redudansi
9. Visualisasi Data

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu mengetahui eksplorasi data. Capaian pembelajaran lulusan pada bab ini adalah adalah adalah capaian pembelajaran pengetahuan adalah mampu menguasai prinsip dan teknik perancangan sistem terintegrasi dengan pendekatan sistem.

3.2. Pengertian Eksplorasi Data

Eksplorasi data terdiri dari 'pembersihan data' normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya. Ada beberapa teknik preprocessing data. Pembersihan data dapat diterapkan untuk

menghilangkan noise dan memperbaiki inkonsistensi dalam data. Integrasi data menggabungkan data dari berbagai sumber ke dalam penyimpanan data yang koheren seperti gudang data. Reduksi data dapat mengurangi ukuran data dengan, misalnya, menggabungkan, menghilangkan fitur yang berlebihan, atau mengelompokkan. Transformasi data (misalnya, normalisasi) dapat diterapkan, di mana data diskalakan agar berada dalam rentang yang lebih kecil seperti 0,0 hingga 1,0. Ini dapat meningkatkan akurasi dan efisiensi algoritma penambangan yang melibatkan pengukuran jarak.

Data dikatakan berkualitas jika memenuhi persyaratan penggunaan yang dimaksudkan. Ada banyak faktor dari kualitas data, termasuk akurasi, kelengkapan, konsistensi, ketepatan waktu, kepercayaan, dan interpretasi.

Data yang tidak akurat, tidak lengkap, dan tidak konsisten adalah properti umum dari basis data dan gudang data dunia nyata yang besar. Ada banyak kemungkinan alasan untuk data yang tidak akurat (yaitu, memiliki nilai atribut yang salah). Instrumen pengumpulan data yang digunakan mungkin salah atau mungkin ada kesalahan manusia atau komputer yang terjadi pada saat memasukkan data. Pada suatu kondisi pengguna dapat dengan sengaja mengirimkan nilai data yang salah untuk bidang wajib ketika mereka tidak ingin mengirimkan informasi pribadi (misalnya, dengan memilih nilai default "1 Januari" yang ditampilkan untuk tanggal kelahiran). Hal ini dikenal sebagai data hilang yang disamarkan. Kesalahan dalam pengiriman data juga dapat terjadi karena ada keterbatasan teknologi seperti ukuran buffer yang terbatas untuk mengoordinasikan transfer dan konsumsi data yang disinkronkan. Data yang salah juga dapat diakibatkan oleh inkonsistensi dalam konvensi penamaan atau kode data, atau format yang tidak konsisten untuk bidang input (misalnya, tanggal). Tuple duplikat juga memerlukan pembersihan data.

Data yang tidak lengkap dapat terjadi karena beberapa alasan. Atribut yang tidak selalu tersedia, seperti informasi pelanggan untuk data transaksi penjualan. Data lain mungkin tidak disertakan hanya karena tidak dianggap penting pada saat diinput. Data yang relevan tidak direkam karena kesalahpahaman atau karena malfungsi peralatan. Data yang tidak konsisten dengan data rekaman lainnya mungkin telah dihapus. Selain itu, pencatatan riwayat data atau modifikasi mungkin diabaikan. Data yang hilang, terutama untuk tupel dengan nilai yang hilang untuk beberapa atribut, mungkin perlu disimpulkan.

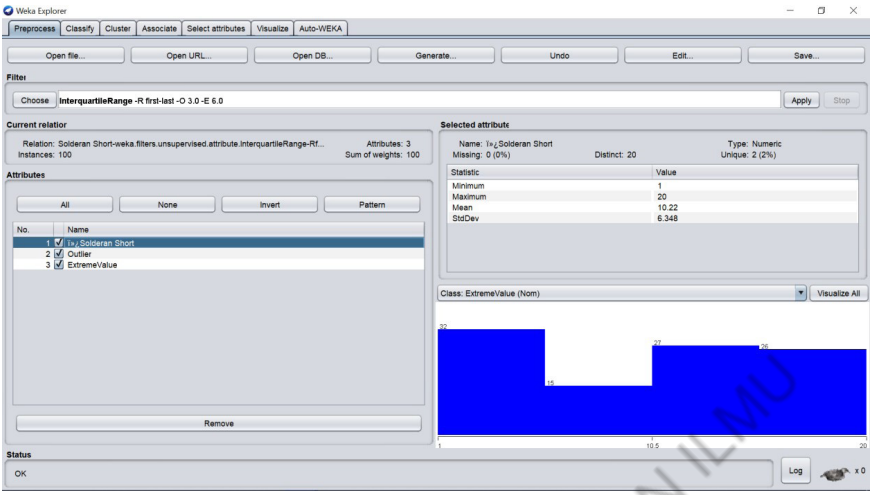
Kualitas data tergantung pada tujuan penggunaan data. Dua pengguna yang berbeda mungkin memiliki penilaian yang sangat berbeda dari kualitas database yang diberikan. Misalnya, seorang analis pemasaran mungkin perlu mengakses database yang disebutkan sebelumnya untuk daftar alamat pelanggan. Beberapa alamat sudah tidak sesuai, namun secara keseluruhan, 80% dari alamat tersebut akurat. Analis pemasaran menganggap ini sebagai basis data pelanggan yang besar untuk tujuan target pemasaran target meskipun, sebagai manajer penjualan.

Ketepatan waktu juga mempengaruhi kualitas data. Dua faktor lain yang mempengaruhi kualitas data adalah kepercayaan dan interpretasi. Kepercayaan data mencerminkan seberapa besar data dipercaya oleh pengguna, sedangkan interpretability mencerminkan seberapa mudah data tersebut dipahami.

3.3 Pembersihan Data

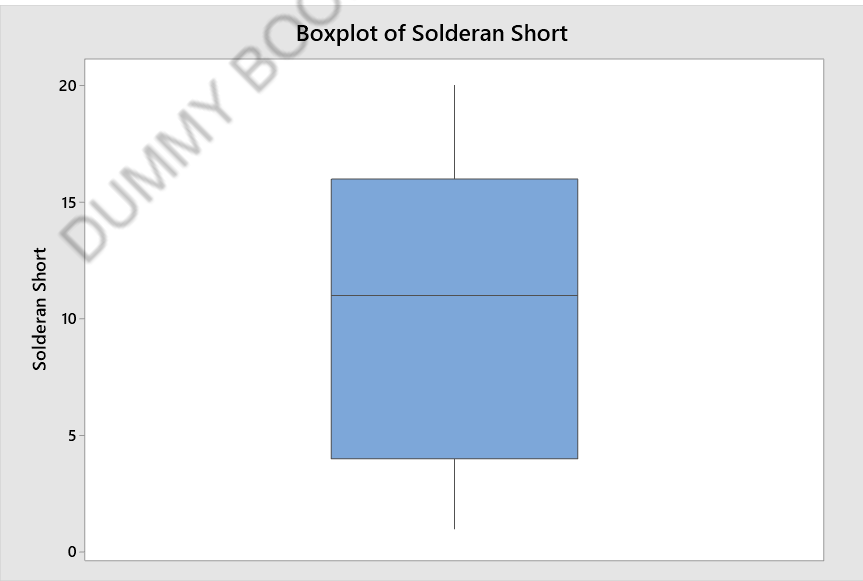
Di dunia nyata data cenderung tidak lengkap, *noisy*, dan tidak konsisten. Pembersihan data rutin dilakukan untuk mengisi nilai yang hilang, menghaluskan noise saat mengidentifikasi outlier, dan memperbaiki inkonsistensi dalam data.

Pada tahapan *data cleaning* dilakukan untuk membersihkan data yang "kotor", maksud dari data "kotor" yaitu data yang menyebabkan informasi *misleading* yang dapat disebabkan oleh adanya duplikasi data, tidak konsistennya data, dan data kosong, hal ini juga dapat disebabkan oleh kesalahan instrumen, kesalahan manusia ataupun kesalahan transmisi. Selanjutnya dilakukan *data reduction* dimana dilakukannya pengurangan data dari kumpulan data, sehingga dengan volume yang jauh lebih kecil dapat menghasilkan hasil analisis yang sama (atau hampir sama), dan dapat dianalisa dengan jangka waktu yang lebih cepat dibandingkan data yang kompleks. Dan pada tahapan terakhir yaitu *data transformation*, tahapan ini berfungsi untuk memetakan seluruh himpunan nilai dari atribut yang diberikan ke himpunan nilai pengganti baru. Dari ketiga tahapan ini akan menghasilkan data yang "bersih" dan siap untuk dimodelkan. Pada data Cleaning digunakan untuk menghilangkan data Outlier dan inkonsentrasi data. Berikut ini merupakan pengolahan data cleaning untuk setiap data jenis kecacatan terhadap produk character display panel (CDP).



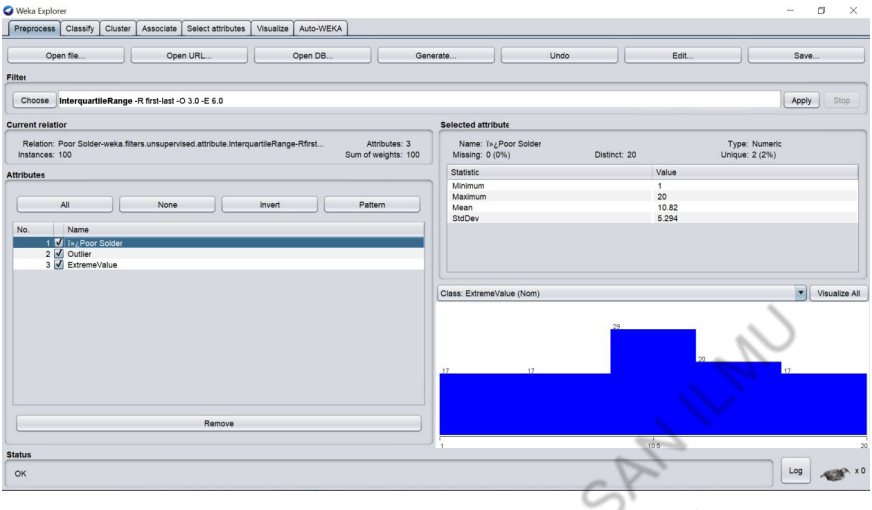
Gambar 3. 1 Pengolahan Data Solderan Short dengan Software WEKA

Pada visualisasi dari hasil Interquartile range dari weka dapat dilihat bahwa data jumlah jenis kecacatan Solderan Short tidak memiliki outlayer, untuk memastikan lebih lanjut dapat menggunakan diagram boxplot melalui *software* MINITAB, berikut ini merupakan hasil pengecekan diagram boxplot dengan menggunakan *software* MINITAB pada data Solderan Short.



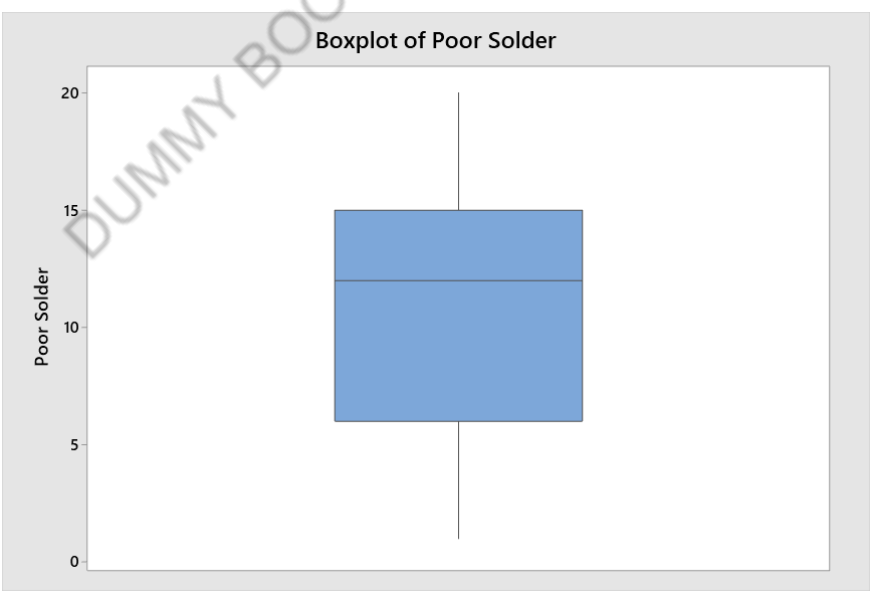
Gambar 3. 2 Grafik Boxplot untuk Data Solderan Short

3.4 Poor Solder



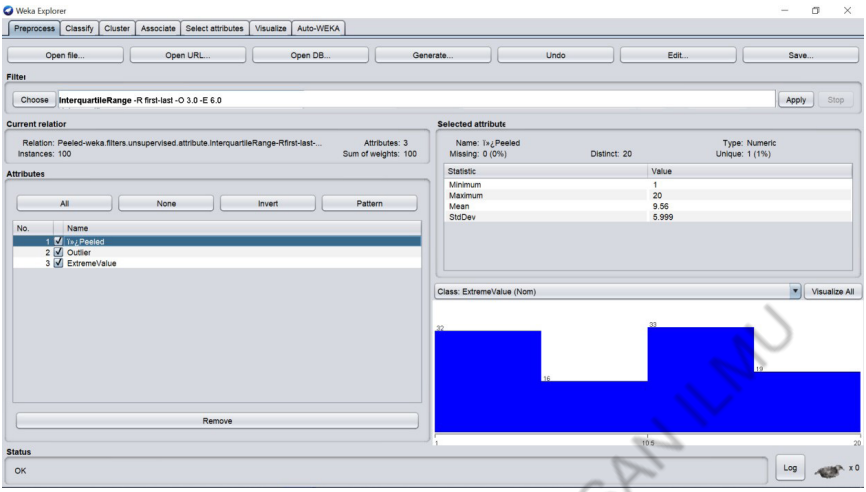
Gambar 3. 3 Pengolahan Data Poor Solder dengan Software WEKA

Pada visualisasi dari hasil Interquartile range dari weka dapat dilihat bahwa data jumlah jenis kecacatan Poor Solder tidak memiliki outlayer, untuk memastikan lebih lanjut dapat menggunakan diagram boxplot melalui *software* MINITAB, berikut ini merupakan hasil pengecekan diagram boxplot dengan menggunakan *software* MINITAB pada data Poor Solder.



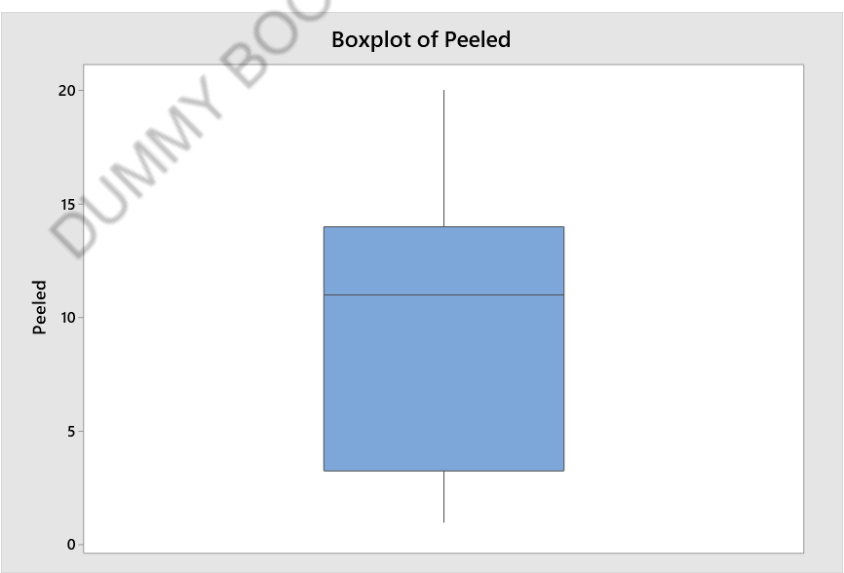
Gambar 3. 4 Grafik Boxplot untuk Data Poor Solder

3.5 Peeled



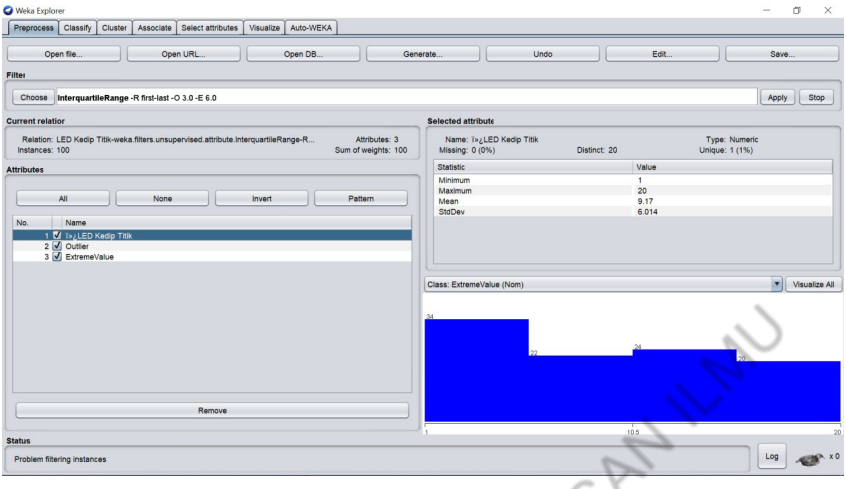
Gambar 3. 5 Pengolahan Data Peeled dengan Software WEKA

Pada visualisasi dari hasil Interquartile range dari weka dapat dilihat bahwa data jumlah jenis kecacatan peeled tidak memiliki outlayer, untuk memastikan lebih lanjut dapat menggunakan diagram boxplot melalui *software* MINITAB, berikut ini merupakan hasil pengecekan diagram boxplot dengan menggunakan *software* MINITAB pada data Peeled.



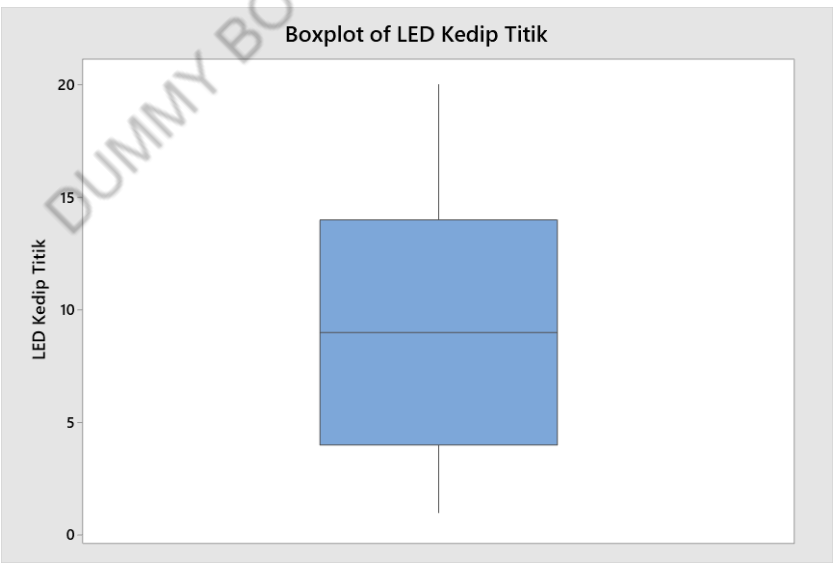
Gambar 3. 6 Grafik Boxplot untuk Data Peeled

3.6 LED Kedip Titik



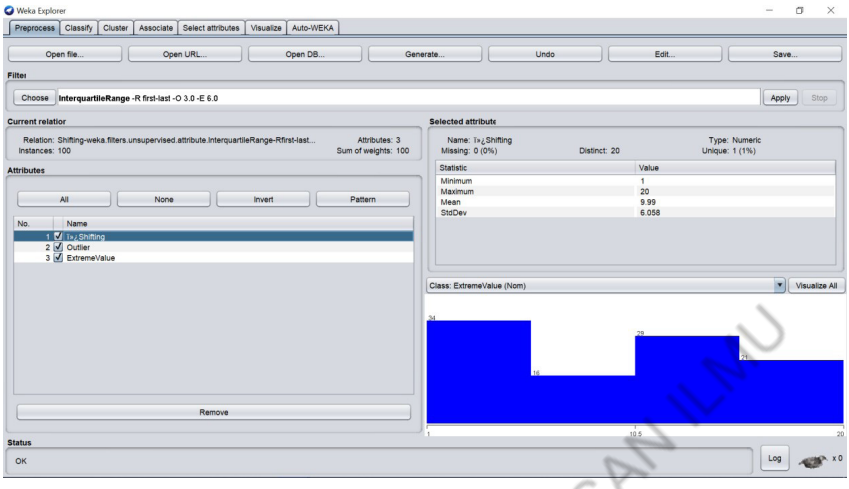
Gambar 3. 7 Pengolahan Data LED Kedip Titik dengan Software WEKA

Pada visualisasi dari hasil Interquartile range dari weka dapat dilihat bahwa data LED Kedip titik tidak memiliki outlayer, untuk memastikan lebih lanjut dapat menggunakan diagram boxplot melalui *software* MINITAB, berikut ini merupakan hasil pengecekan diagram boxplot dengan menggunakan *software* MINITAB pada data LED Kedip Titik.



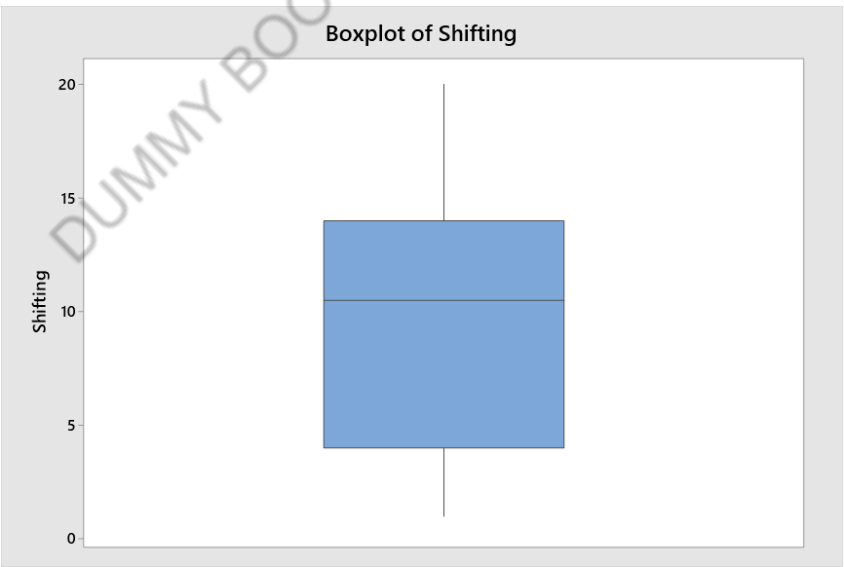
Gambar 3. 8 Grafik Boxplot untuk Data LED Kedip Titik

3.7 Shifting



Gambar 3. 9 Pengolahan Data LED Kedip Titik dengan Software WEKA

Pada visualisasi dari hasil Interquartile range dari weka dapat dilihat bahwa data jumlah jenis kecacatan Shifting tidak memiliki outlayer, untuk memastikan lebih lanjut dapat menggunakan diagram boxplot melalui *software* MINITAB, berikut ini merupakan hasil pengecekan diagram boxplot dengan menggunakan *software* MINITAB pada data Shifting.



Gambar 3. 10 Grafik Boxplot untuk Data Shifting

	A	B	C	D	E
1	lokasi_omzet_penjualan	jenis_komoditi	volume	satuan	omzet_(rp)
2	Pasar Bunga Rawabelong	Melati	8221	Bks	283975000
3	Pasar Bunga Rawabelong	Pihong	1847	Ktg	21525000
4	Pasar Bunga Rawabelong	Kenanga	905	Ktg	18395000
5	Pasar Bunga Rawabelong	Kantil	20	Bks	1000000
6	Pasar Bunga Rawabelong	Ros Kampung/Ros Cipanas	0	Ikat	0
7	Pasar Bunga Rawabelong	Ros Rontokan Malang	1303	Ktg	33820000
8	Pasar Bunga Rawabelong	Mawar Tabur	1028	Bks	17065000
9	Pasar Bunga Rawabelong	Pandan Iris	1808	Ktg	21080000
10	Pasar Bunga Rawabelong	Daun Sirih	325	Ikat	1625000
11	Pasar Bunga Rawabelong	Buah Pinang	245	Ikat	2450000
12	Pasar Bunga Rawabelong	Nanas Muda	45	Buah	675000
13	Pasar Bunga Rawabelong	Sedap Malam	750	Gabung	150000000
14	Pasar Bunga Rawabelong	Gladiol	336	Ikat	8400000
15	Pasar Bunga Rawabelong	Ros Semy Holand Cipanas	2844	Ikat	227520000
16	Pasar Bunga Rawabelong	Ros Malang	3235	Ikat	144280000
17	Pasar Bunga Rawabelong	Hebras/Gerbera	1103	Ikat	66990000
18	Pasar Bunga Rawabelong	Lely	27	kuntum	1350000
19	Pasar Bunga Rawabelong	Pikok	4980	Ikat	82795000
20	Pasar Bunga Rawabelong	Bunga Balon	11	Ikat	110000
21	Pasar Bunga Rawabelong	Aster (Cipanas)	1196	Ikat	19875000
22	Pasar Bunga Rawabelong	Krisan (Cipanas)	984	Ikat	21300000
23	Pasar Bunga Rawabelong	Bunga Matahari	176	Ikat	8620000
24	Pasar Bunga Rawabelong	Bunga Teratai	0	Ikat	0
25	Pasar Bunga Rawabelong	Aster (PT)	6720	Ikat	111440000
26	Pasar Bunga Rawabelong	Krisan Std (PT)	2964	Ikat	64170000
27	Pasar Bunga Rawabelong	Gerbera (PT)	1606	Ikat	32120000
28	Pasar Bunga Rawabelong	Casablanca	1015	Ikat	172810000

06-2021 Sheet1

Gambar 3.11 Dataset berisikan tentang data omzet pemasaran bunga dan tanaman hias di Provinsi DKI Jakarta Bulan Januari s.d. Juni Tahun 2021.

Keterangan variabel data ;

lokasi_omset_penjualan : Nama Lokasi Penjualan jenis_ komoditi : Jenis penjualan bunga dan tanaman hias.

Volume : Jumlah volume jenis komoditi Satuan : Satuan jenis komoditi omzet_(rp) : Hasil pendapatan penjualan (dalam rupiah) Dataset berisikan tentang data omzet pemasaran bung) : https:// katalog.data.go.id/dataset/data-jumlah- omzet-pemasaran-bunga-dan-tanaman-hias-di- provinsi-dki-jakarta-2021

Data pre processing dilakukan dengan menggunakan Rapid Miner.

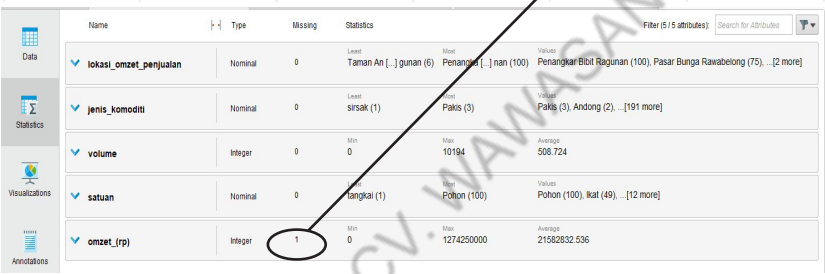
3.8 Nilai yang Hilang

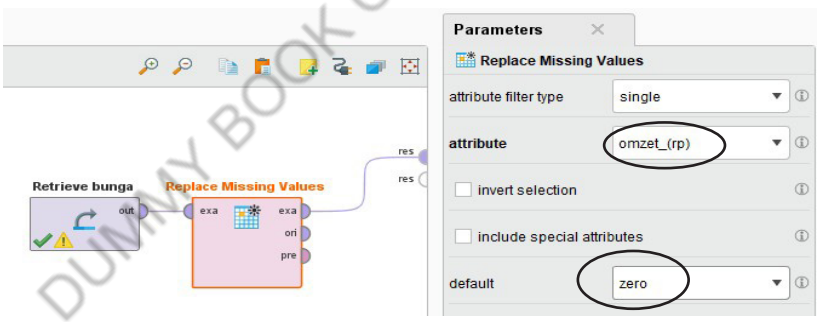
Jika terdapat suatu data yang hilang (*missing value*), maka untuk mengatasi hal ini dapat dilakukan beberapa metode, diantaranya adalah:

1. **Abaikan Tuple (*Ignore the tupple*)**: Metode ini biasanya dilakukan ketika label kelas hilang (dengan asumsi data mining yang dilakukan menggunakan klasifikasi). Metode ini tidak terlalu efektif, kecuali jika tuple berisi beberapa atribut dengan nilai yang hilang. Metode ini tidak disarankan ketika persentase nilai yang hilang per atribut sangat bervariasi.
2. **Mengisi nilai yang hilang secara manual**: Secara umum, pendekatan ini memakan waktu dan tidak dapat dilakukan pada kumpulan data yang memiliki nilai hilang yang banyak.
3. **Gunakan konstanta global untuk mengisi nilai yang hilang**: Ganti semua nilai atribut yang hilang dengan konstanta yang sama seperti label seperti "Tidak diketahui".
4. **Gunakan ukuran tendensi sentral untuk atribut** (misalnya mean atau median) untuk mengisi nilai yang hilang.
5. **Gunakan atribut mean atau median untuk semua sampel yang termasuk dalam kelas yang sama dengan tupel yang diberikan**: Misalnya, jika mengklasifikasikan pelanggan menurut risiko kredit, maka dapat mengganti nilai yang hilang dengan nilai pendapatan rata-rata untuk pelanggan dalam risiko kredit yang sama kategori sebagai tupel yang diberikan. Jika distribusi data untuk kelas tertentu miring, nilai median adalah pilihan yang lebih baik.
6. **Gunakan nilai yang paling mungkin untuk mengisi nilai yang hilang**: Ini dapat ditentukan dengan regresi, alat berbasis inferensi menggunakan formalisme Bayesian, atau pohon keputusan. Misalnya, menggunakan atribut pelanggan lain dalam kumpulan data, maka dapat membuat pohon keputusan untuk memprediksi nilai pendapatan yang hilang.

Metode 3 sampai 6 membiaskan data—nilai yang diisi mungkin tidak benar. Metode 6 adalah strategi yang populer. Metode ini menggunakan sebagian besar informasi dari data saat ini untuk memprediksi nilai yang hilang. Dengan mempertimbangkan nilai atribut lain dalam estimasi nilai yang hilang untuk pendapatan, ada kemungkinan lebih besar bahwa hubungan antara pendapatan dan atribut lainnya dipertahankan.

Row No.	lokasi_omz...	jenis_komo...	volume	satuan	omzet_(r...
110	Taman Anggr...	Tanaman Hia...	1331	Pot	?
105	Taman Anggr...	Phalaenopsis	10194	Pot	1274250000
44	Pasar Bunga ...	Anggrek Bulan	2753	Pot	344125000
1	Pasar Bunga ...	Melati	8221	Bks	283975000
14	Pasar Bunga ...	Ros Semy H...	2844	Ikat	227520000





Row No.	lokasi_omz...	jenis_komo...	volume	satuan	omzet_(rp)
110	Taman Anggr...	Tanaman Hia...	1331	Pot	0
111	Penangkar Bi...	Pohon sukun	3	Pohon	750000
112	Penangkar Bi...	Nangka Cem...	0	Pohon	0
113	Penangkar Bi...	Nangka Mini	8	Pohon	1200000
114	Penangkar Bi...	Apel Miki	0	Pohon	0

Gambar 3.12 Pembersihan Data dengan menggunakan software Rapid Miner

Pada beberapa kasus, nilai yang hilang mungkin tidak menyiratkan kesalahan dalam data. Misalnya, saat mengajukan permohonan kartu kredit, kandidat mungkin diminta untuk memberikan nomor SIM mereka. Kandidat yang tidak memiliki SIM dapat dengan sendirinya mengosongkan bidang ini. Formulir harus memungkinkan responden untuk menentukan nilai seperti "tidak berlaku." Rutinitas perangkat lunak juga dapat digunakan untuk mengungkap nilai null lainnya (misalnya, "tidak tahu", "?", atau "tidak ada"). Idealnya, setiap atribut harus memiliki satu atau lebih aturan mengenai kondisi nol. Aturan dapat menentukan apakah nol diperbolehkan atau tidak dan/atau bagaimana nilai tersebut harus ditangani atau diubah. Kolom mungkin juga sengaja dikosongkan jika akan disediakan di langkah selanjutnya dari proses bisnis. Oleh karena itu, meskipun kita dapat mencoba yang terbaik untuk membersihkan data setelah disita, database yang baik dan desain prosedur entri data akan membantu meminimalkan jumlah nilai yang hilang atau kesalahan sejak awal.

3.9 Noisy Data

Noise adalah kesalahan acak atau variasi-variasi pada data dalam variable yang diukur. Dalam Bab 2, telah dijelaskan beberapa teknik dasar statistic deskriptif (misalnya, boxplot dan plot pencar), dan metode visualisasi data yang dapat digunakan untuk mengidentifikasi outlier, yang mungkin mewakili noise.

Binning: Metode binning dilakukan dengan memerikasa "nilai tetangga" yang artinya adalah nilai-nilai yang terbesar. Kemudian data yang telah diurutkan di partisi dalam beberapa bin. Selanjutnya akan dilakukan **smoothing by bin means, smoothing by bin medians,** dan **smoothing by bin boundaries.**

Dalam **smoothing by bin mean**s, setiap nilai dalam bin diganti dengan nilai rata-rata bin. Misalnya, rata-rata nilai 4, 8, dan 15 di Bin 1 adalah 9. Oleh karena itu, setiap nilai asli di bin ini diganti dengan nilai 9.

Demikian pula, **smoothing by bin median** dapat digunakan, di mana setiap nilai bin diganti dengan median bin. Pada **smoothing by bin boundaries**, nilai minimum dan maksimum dalam bin yang diberikan diidentifikasi sebagai batas bin. Setiap nilai bin kemudian diganti dengan nilai batas terdekat. Secara umum, semakin besar lebarnya, semakin besar efek smoothingnya.

Regresi: Data smoothing juga dapat dilakukan dengan regresi, yaitu suatu teknik yang menyesuaikan nilai data dengan suatu fungsi. Regresi linier melibatkan pencarian garis "terbaik" agar sesuai dengan dua atribut (atau variabel) sehingga satu atribut dapat digunakan untuk memprediksi yang lain. Regresi linier berganda adalah perpanjangan dari regresi linier, di mana lebih dari dua atribut terlibat dan data sesuai dengan permukaan multidimensi.

Analisis outlier: Pencilan dapat dideteksi dengan pengelompokan, misalnya, di mana nilai-nilai serupa diatur ke dalam kelompok, atau "cluster."

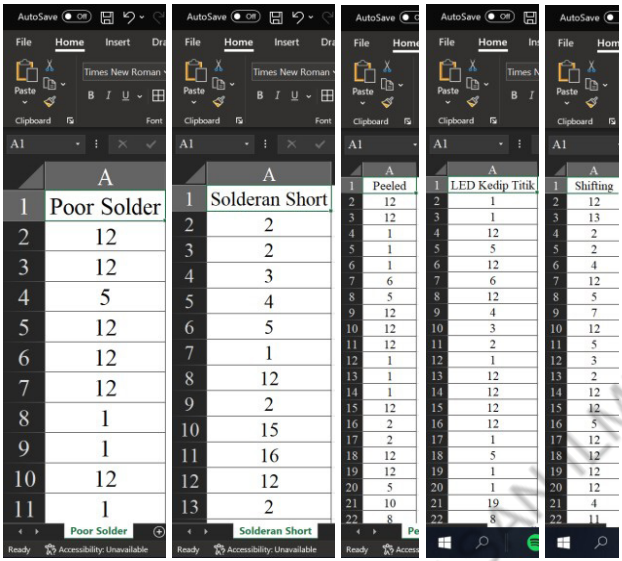
Banyak metode data smoothing juga digunakan untuk diskritisasi data (suatu bentuk transformasi data) dan reduksi data. Misalnya, teknik binning yang dijelaskan sebelumnya mengurangi jumlah nilai yang berbeda per atribut. Ini bertindak sebagai bentuk reduksi data untuk metode data mining berbasis logika, seperti induksi pohon keputusan, yang berulang kali membuat perbandingan nilai pada data yang diurutkan.

Konsep hirarki adalah bentuk diskritisasi data yang juga dapat digunakan untuk data smoothing. Konsep hirarki untuk harga, misalnya, dapat memetakan nilai harga riil menjadi murah, harga sedang, dan mahal, sehingga mengurangi jumlah nilai data yang harus ditangani oleh data mining.

3.10. Data Integrasi

Data mining seringkali membutuhkan integrasi data—penggabungan data dari beberapa penyimpanan data. Integrasi yang cermat dapat membantu mengurangi dan menghindari redundansi dan inkonsistensi dalam kumpulan data yang dihasilkan. Hal ini dapat membantu meningkatkan akurasi dan kecepatan proses data mining selanjutnya.

Data Integration atau integrasi data merupakan proses menggabungkan atau menyatukan dua atau lebih sebuah data dari berbagai sumber database yang berbeda ke dalam sebuah penyimpanan seperti gudang data (data warehouse). Pada integrasi data ini dari kelima jenis kecacatan disatukan kedalam satu file excel dengan format xlsx. Berikut ini merupakan tahapan integrasi data yang dilakukan.



	A
1	Poor Solder
2	12
3	12
4	5
5	12
6	12
7	12
8	1
9	1
10	12
11	1

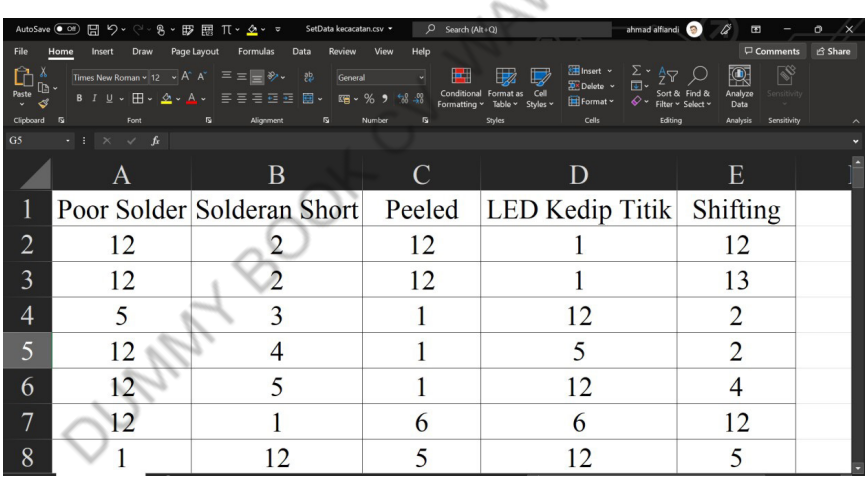
	A
1	Solderan Short
2	2
3	2
4	3
5	4
6	5
7	1
8	12
9	2
10	15
11	16
12	12
13	2

	A
1	Peeled
2	12
3	12
4	1
5	1
6	1
7	6
8	5
9	12
10	12
11	12
12	1
13	1
14	1
15	12
16	2
17	2
18	12
19	12
20	5
21	10
22	8

	A
1	LED Kedip Titik
2	1
3	1
4	12
5	5
6	12
7	6
8	12
9	4
10	3
11	2
12	1
13	12
14	12
15	12
16	12
17	1
18	5
19	1
20	1
21	19
22	8

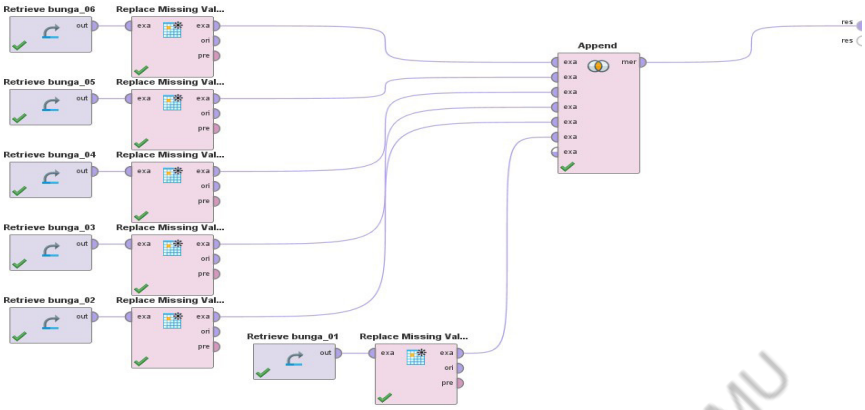
	A
1	Shifting
2	12
3	13
4	2
5	2
6	4
7	12
8	5
9	7
10	12
11	5
12	3
13	2
14	12
15	12
16	5
17	12
18	12
19	12
20	12
21	4
22	11

Gambar 3. 13 Data Kelima Jenis Kecacatan



	A	B	C	D	E
1	Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting
2	12	2	12	1	12
3	12	2	12	1	13
4	5	3	1	12	2
5	12	4	1	5	2
6	12	5	1	12	4
7	12	1	6	6	12
8	1	12	5	12	5

Gambar 3. 14 Set Data Setelah Data Integration



Gambar 3. 15 Integrasi Data

Dataset omzet pemasaran bunga bulan Januari s.d. Juni 2021 yang terpecah menjadi 6 single file berformat CSV digabung menjadi satu dataset

Tabel 3. 1 Integrasi Data menggunakan Software Knime

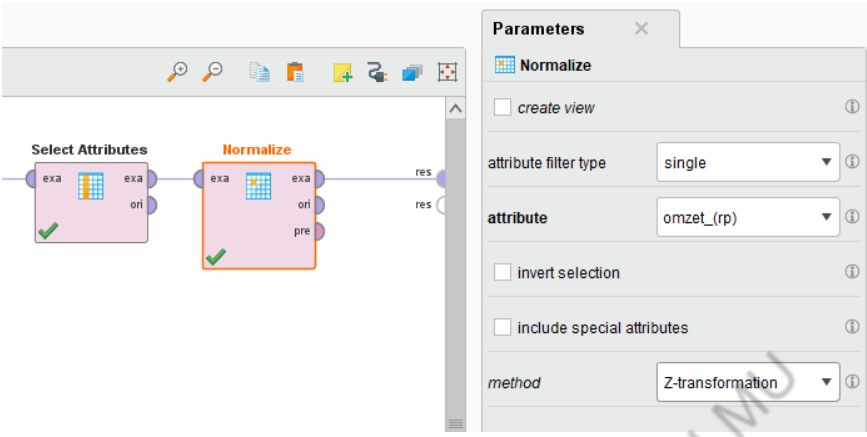
Row No.	omzet_(rp)	lokasi_omz...	jenis_komo...	volume	satuan
1	283975000	Pasar Bunga ...	Melati	8221	Bks
2	21525000	Pasar Bunga ...	Pihong	1847	Ktg
3	18395000	Pasar Bunga ...	Kenanga	905	Ktg
4	1000000	Pasar Bunga ...	Kantil	20	Bks
5	0	Pasar Bunga ...	Ros Kampun...	0	Ikat
6	33820000	Pasar Bunga ...	Ros Rontoka...	1303	Ktg
7	17065000	Pasar Bunga ...	Mawar Tabur	1028	Bks
8	21080000	Pasar Bunga ...	Pandan Iris	1808	Ktg
9	1625000	Pasar Bunga ...	Daun Sirih	325	Ikat
10	2450000	Pasar Bunga ...	Buah Pinang	245	Ikat
11	675000	Pasar Bunga ...	Nanas Muda	45	Buah
12	150000000	Pasar Bunga ...	Sedap Malam	750	Gabung
13	8400000	Pasar Bunga ...	Gladiol	336	Ikat
14	227520000	Pasar Bunga ...	Ros Semy H...	2844	Ikat
15	144280000	Pasar Bunga ...	Ros Malang	3235	Ikat
16	66990000	Pasar Bunga ...	Hebras/Gerb...	1103	Ikat
17	1350000	Pasar Bunga ...	Lely	27	kuntum
18	82795000	Pasar Bunga ...	Pikok	4980	Ikat
19	110000	Pasar Bunga ...	Bunga Balon	11	Ikat

3.11. Transformasi Data

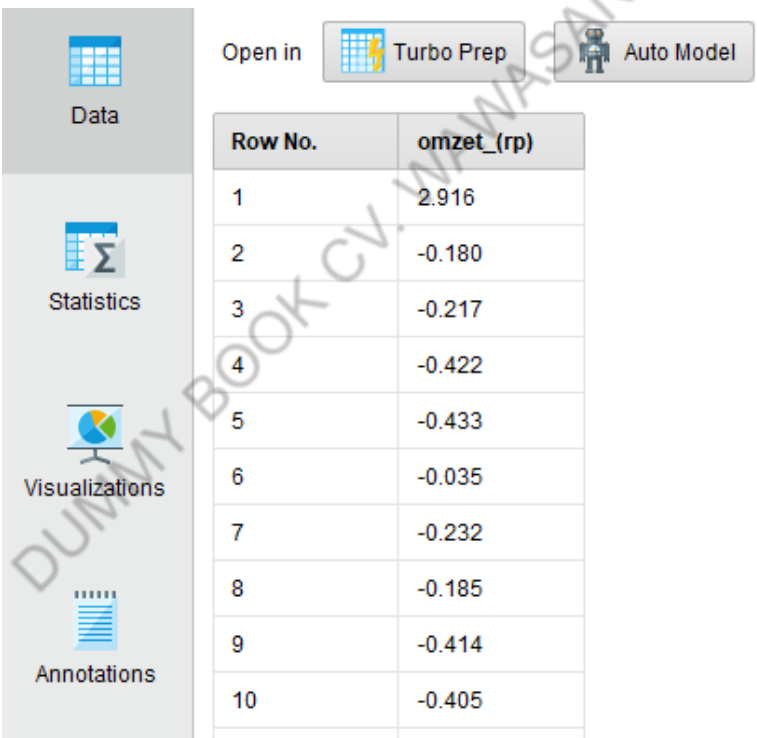
Data Transformation adalah tahapan untuk dapat mengubah bentuk data yang dilakukan agar set data siap diproses untuk dapat menghasilkan analisis yang lebih baik. Bentuk data kecacatan produk yang diperoleh merupakan data numerik, berupa jumlah produk *reject* pada tiap jenis kecacatan. Tiap atribut jenis kecacatan ditransformasikan menjadi "YES" apabila terdapat kecacatan tinggi diatas 5 kecacatan dan "NO" apabila terdapat kecacatan kurang dari 5 kecacatan. Transformasi data ini dapat menggunakan formula =IF(A2<5;"NO";"YES") pada *software microsoft excel*. Berikut ini merupakan contoh hasil dari transformasi data.

Tabel 3. 2 Set data Hasil Transformasi

Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting
YES	NO	YES	NO	YES
YES	NO	YES	NO	YES
YES	NO	NO	YES	NO
YES	NO	NO	YES	NO
YES	YES	NO	YES	NO
YES	NO	YES	YES	YES
NO	YES	YES	YES	YES
NO	NO	YES	NO	YES
YES	YES	YES	NO	YES
NO	YES	YES	NO	YES
YES	YES	NO	NO	NO
YES	NO	NO	YES	NO
YES	YES	NO	YES	YES
YES	NO	YES	YES	YES



Gambar 3. 16 Transformasi data menggunakan Rapid Miner (1)



Gambar 3. 17 Transformasi Data menggunakan Rapid Miner (2)

Transformasi data dengan normalisasi (proses penskalaan nilai atribut dari data pada suatu range tertentu).

3.12. Reduksi Data

Fase selanjutnya adalah *data reduction* di mana dilakukannya pengurangan pada set data yang ada. Pengurangan ini didasari dengan beberapa lot inspeksi yang tidak memiliki jenis kecacatan, sehingga produknya termasuk kedalam *acceptable goods*. Sehingga untuk data dengan persentase *reject* 0% tidak diperhitungkan, karena pembahasan data difokuskan pada klasifikasi *reject* tinggi dan *reject* rendah. Dengan menggunakan data *reduction* ini, mengakibatkan berkurangnya data yang pada sampel 48 hingga 55 data. Sehingga data yang ada berkurang yang sebelumnya berjumlah 100 menjadi 93.

Tabel 3. 3 Contoh Pola Data No Pada Kelima Jenis Atribut Jenis Kecacatan

Sampel	Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting
48	NO	NO	NO	NO	NO
49	NO	NO	NO	NO	NO
50	NO	NO	NO	NO	NO
51	NO	NO	NO	NO	NO
52	NO	NO	NO	NO	NO
53	NO	NO	NO	NO	NO
54	NO	NO	NO	NO	NO
55	NO	NO	NO	NO	NO
56	NO	NO	YES	YES	NO
57	NO	NO	YES	YES	YES
58	YES	YES	YES	YES	YES
59	YES	YES	YES	YES	NO
60	YES	NO	NO	YES	YES

3.13. Analisis Korelasi dan Redudansi

Redudansi adalah masalah penting dalam integrasi data. Atribut (seperti pendapatan tahunan, misalnya) mungkin berlebihan jika dapat "diturunkan" dari atribut atau kumpulan atribut lain. Inkonsistensi dalam penamaan atribut atau dimensi juga dapat menyebabkan redundansi dalam kumpulan data yang

dihasilkan. Beberapa redudansi dapat dideteksi dengan analisis korelasi. Mengingat dua atribut, analisis tersebut dapat mengukur seberapa kuat satu atribut menyiratkan yang lain, berdasarkan data yang tersedia.

Untuk data nominal, hubungan korelasi antara dua atribut, A dan B, dapat ditemukan dengan X^2(chi square). Misalkan A memiliki nilai yang berbeda, yaitu a1,a2,...ac. B memiliki r nilai yang berbeda, yaitu b1,b2,...br. Tupel data yang dijelaskan oleh A dan B dapat ditampilkan sebagai tabel kontingensi, dengan nilai c dari A membentuk kolom dan nilai r dari B membentuk baris. Misalkan (Ai , Bj) menyatakan kejadian bersama bahwa atribut A mengambil nilai ai dan atribut B mengambil nilai bj, yaitu, di mana (A = ai,B = bj). Setiap kejadian gabungan (Ai,Bj) yang mungkin memiliki sel (atau slot) sendiri dalam tabel. Nilai X^2 dapat dihitung dengan

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Statistik X^2 menguji hipotesis bahwa A dan B independen, yaitu tidak ada korelasi di antara keduanya. Pengujian didasarkan pada tingkat signifikansi, dengan (r-1) × (c-1) derajat kebebasan. Jika hipotesis nol ditolak, maka dapat dikatakan bahwa A dan B berkorelasi secara statistik.

Contoh:

Misalkan sekelompok 1500 orang disurvei. Jenis kelamin setiap orang dicatat. Setiap orang disurvei, apakah jenis bahan bacaan yang disukainya adalah fiksi atau nonfiksi. Jadi, terdapat dua atribut, jenis kelamin dan bacaan yang disukai. Hasil pengamatan ditunjukkan pada Tabel 3.4 berikut

Tabel 3. 4 Hasil dari pengamatan survei

	Pria	Wanita	Total
Fiksi	250	200	450
Non fiksi	50	1000	1050
Total	300	1200	1500

Misalnya:

Untuk perhitungan ekspektasi frekuensi pada pria adalah sebagai berikut:

$$e_{11} = \frac{\text{Jumlah pria} \times \text{Jumlah fiksi}}{n} = \frac{300 \times 450}{1500} = 90$$

Perhitungan tersebut dilanjutkan sehingga menghasilkan nilai pada Tabel 3.5.

Tabel 3. 5 Hasil perhitungan ekspektasi frekuensi

	Pria	**Wanita**	**Total**
Fiksi	250 (90)	200 (360)	450
Non fiksi	50 (210)	1000 (840)	1050
Total	300	1200	1500

Untuk X^2 (chi square):

$$X^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{90}$$

$$= 507{,}93$$

Untuk tabel 2 × 2 ini, derajat kebebasannya adalah (2-1) (2-1) = 1. Untuk 1 derajat kebebasan, nilai X^2 yang diperlukan untuk menolak hipotesis pada tingkat signifikansi 0,001 adalah 10,828 (diambil dari tabel poin persentase atas dari distribusi chi square). Karena nilai yang dihitung di atas ini, maka hipotesis ditolak bahwa jenis kelamin dan bacaan yang disukai adalah independen dan menyimpulkan bahwa kedua atribut tersebut (sangat) berkorelasi untuk kelompok orang tertentu.

3.14. Visualisasi

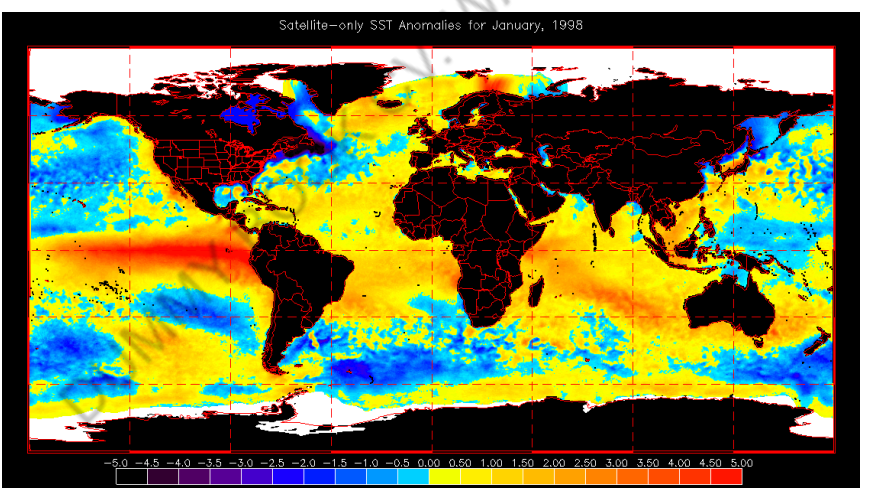
Visualisasi data adalah menampilkan informasi dalam format grafik atau tabular. Visualisasi yang baik memerlukan data (informasi) dikonversi ke dalam format visual sedemikian sehingga karakteristik dari data dan hubungan diantara item data atau atribut dapat dianalisa. Teknik visualisasi dalam data mining dinyatakan sebagai **visual data mining**.

Motivasi untuk menggunaan visualisasi adalah bahwa pengguna dapat dengan cepat menyerap sejumlah besar informasi visual dan menemukan pola dalam informasi tersebut. Perhatikan Gambar 3.2, yang menunjukkan Sea Surface Termperature (SST) dalam derajat Celcius untuk klimatologi didasarkan pada pengamatan malam hari untuk Januari 1998.

Produk Anomali Suhu Permukaan Laut (SST) 5km global ini, menampilkan perbedaan antara SST saat ini dan rata-rata jangka panjang. Skala berubah dari -5 hingga +5 °C. Angka positif berarti suhu lebih hangat dari rata-rata; negatif berarti lebih dingin dari rata-rata.

Bidang anomali SST (derajat C) adalah perbedaan antara SST hanya 50 km malam hari dan klimatologi SST rata-rata bulanan hanya malam hari.

Gambar tersebut meringkas informasi dari sekitar 250.000 angka dan dapat dengan mudah diinterpretasikan dalam beberapa detik. Sebagai contoh, dapat dilihat dengan mudah temperatur laut adalah paling tinggi pada garis khatulistiwa dan paling rendah di kutub.



Gambar 3.18 Satelit untuk SST Januari 1998 (www.ospo.noaa.gov)

Motivasi lainnya untuk visualisasi adalah membuat penggunaan domain knowledge. Walaupun penggunaan domain knowledge adalah pekerjaan yang penting dalam data mining, seringkali sulit dan tidak mungkin mengunakan seluruh pengetahuan tersebut dalam alat statistik atau algoritmik. Dalam beberapa kasus, analisa

dapat dilakukan dengan menggunakan alat non-visual dan kemudian hasilnya dipresentasikan secara visual untuk dievaluasi oleh domain expert. Dalam kasus lain, setelah menemukan pola yang diinginkan, karena dengan menggunakan domain knowledge, pengguna dapat dengan cepat membuang beberapa pola yang tidak menarik dan langsung terfokus pada pola yang penting.

Berikut adalah pendekatan-pendekatan yang umum untuk visualisasi data dan atributnya.

a. Representasi: Pemetaan Data ke Elemen Grafik

Langkah pertama dalam visualisasi adalah pemetaan informasi ke format visual; yaitu pemetaan objek, atribut, dan hubungan antar objek dalam sekumpulan informasi ke objek, atribut dan hubungan visual. Bahwa objek data, atributnya, dan hubungan antar objek data dinyatakan dalam elemen-elemen grafis seperti titik, garis, bentuk dan warna. Objek biasanya direpresentasikan dalam salah satu dari tiga cara berikut. Pertama, jika hanya sebuah atribut kategori dari objek yang diperhatikan, maka objek seringkali disatukan ke dalam kategori-kategori berdasarkan pada nilai atribut tersebut, dan kategori-kategori ini ditampilkan sebagai sebuah entri dalam tabel atau area di layar. Kedua, jika sebuah objek memiliki banyak atribut, maka objek dapat ditampilkan sebagai baris (atau kolom) dari sebuah tabel atau sebuah garis pada sebuah grafik. Ketiga, sebuah objek serngkali diinterpretasikan sebagai sebuah titik dalam ruang 2 atau 3 dimensi, dimana secara grafis, titik dapat direpresentasikan oleh gambar geometri seperti lingkaran dan kotak. Untuk atribut, representasi tergantung pada tipe atribut, apakah nominal, ordinal, atau kontinu (interval atau ratio). Atribut ordinal dan kontinu dapat dipetakan ke dalam fitur grafis terurut dan kontinu seperti lokasi sepanjang sumbu x, y dan z; intensitas; warna; atau ukuran (diameter, tinggi dan lain-lain). Untuk atribut kategori, setiap kategori dapat dipetakan ke dalam posisi, warna, bentuk, orientasi yang berbeda atau kolom dalam tabel. Untuk atribut nominal, yang memiliki nilai terurut, penggunaan fitur-fitur grafik, seperti warna dan posisi yang memiliki urutan terkait dengan nilai-nilainya, harus dilakukan secara hati-hati. Representasi hubungan melalui elemen-elemen grafis terjadi baik secara eksplisit maupun implisit. Untuk data graf, digunakan representasi graf biasa, sekumpulan node dengan link diantara node. Jika node (objek data) atau link (hubungan) memiliki atribut atau karakteristik dari dirinya sendiri, maka atribut dan karakteristik tersebut direpresentasikan secara

grafis. Sebagai ilustrasi, jika node adalah kota dan link adalah jalan raya, maka diameter dari node dapat menyatakan populasi, sedangkan lebar dari link dapat merepresentasikan volume lalu lintas.

Dalam banyak kasus, pemetaan objek dan atribut ke elemen grafis secara implisit memetakan hubungan dalam data ke hubungan antara elemen-elemen grafis. Sebagai ilustrasi, jika objek data merepresentasikan objek fisik yang memiliki lokasi, seperti kota, maka posisi relatif dari objek grafis yang berhubungan dengan objek data cenderung mempertahankan posisi relatif aktual dari data.

b. Penyusunan

Pemilihan yang tepat dari representasi visual dari objek dan atribut adalah penting untuk visualisasi yang baik. Penyusunan kembali item dalam penampilan visual juga merupakan hal yang penting.

c. Seleksi

Konsep penting lainnya dalam visualisasi adalah seleksi, yang mengeliminasi objek atau atribut tertentu. Jika terlalu banyak objek data, maka memvisualisasikan semua objek akan menghasilkan tampilan yang penuh sesak. Pendekatan yang paling umum untuk menangani atribut yang banyak adalah dengan memilih sebuah subset dari atribut. Jika dimensi terlalu tinggi, matriks plot untuk dua atribut dapat dibuat untuk menggambarkan objek data secara simultan.

Teknik memilih sepasang (atau sejumlah kecil) atribut adalah bentuk dari reduksi dimensionalitas, dan terdapat banyak teknik yang dapat digunakan, salah satunya adalah PCA (Principal Components Analysis).

3.3.3. Metode Visualisasi

Teknik visualisasi seringkali ditentukan berdasarkan tipe dari data yang sedang dianalisis, berdasarkan banyaknya atribut yang terlibat, berdasarkan tipe atribut atau berdasarkan karakteristik khusus dari data seperti struktur hirarki atau graf.

Visualisasi Sejumlah Kecil Atribut

Terdapat beberapa teknik yang dapat digunakan untuk visualisasi data dengan jumlah atribut yang sedikit. Beberapa

teknik tersebut, seperti histogram, memberikan distribusi nilai yang diobservasi untuk satu atribut. Sedangkan teknik yang lain seperti scatter plot digunakan untuk menampilkan hubungan antara nilai dari dua atribut.

Stem and Leaf Plot.

Stem and leaf plot dapat digunakan untuk mendapatkan distribusi dari data integer atau kontinu satu dimensi. Untuk bentuk sederhana dari *stem and leaf plot*, kita bagi nilai-nilai ke dalam dua grup, dimana setiap grup mengandung nilai-nilai yang sama kecuali untuk digit terakhirnya. Dengan demikian, jika nilai-nilai tersebut adalah integer dua digit, contoh 35, 36, 42, dan 51, maka stem adalah digit pada ordo tertinggi, yaitu 3, 4, 5, sedangkan leaf adalah digit dengan ordo rendah, yaitu 1, 2, 5, dan 6. Dengan memplotkan stem secara vertikal dan leaf secara horizontal, maka dapat diperoleh representasi visual dari distribusi data.

3.15. Latihan Soal

1. Jelaskan pre processing data dari data pada jurnal dengan judul "Exploring feature selection and classification methods for predicting heart disease " pada link Exploring feature selection and classification methods for predicting heart disease (sagepub.com) (https://journals.sagepub.com/doi/pdf/10.1177/2055207620914777)
2. Jelaskan konsep Feature Selection

BAB IV
KLASIFIKASI DAN DECISION TREE

4.1. Tujuan dan Capaian Pembelajaran

Tujuan:

Mahasiswa mampu melakukan investigasi, menganalisa, menginterpretasikan data dan informasi dan membuat keputusan yang tepat berdasarkan pendekatan analisa klasifikasi. Bab ini bertujuan untuk mengenalkan dasar klasifikasi dan *decision tree* (pohon keputusan).

Materi yang diberikan :

1. Klasifikasi
2. Decision Tree/ Pohon Keputusan

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu melakukan investigasi, menganalisa, menginterpretasikan data dan informasi dan membuat keputusan yang tepat, menguasai konsep dan aplikasi berdasarkan pendekatan analisa klasifikasi dan *decision tree*. Capaian pembelajaran lulusan pada bab ini adalah capaian pembelajaran pengetahuan adalah mampu menguasai prinsip dan teknik perancangan sistem terintegrasi dengan pendekatan sistem.

4.2. Klasifikasi

Klasifikasi dapat dideskripsikan sebagai algoritma dalam proses pembelajaran mesin. Hal ini dapat diberikan label ke dalam objek data berdasarkan pengetahuan dari kelas dimana data di simpan. Dalam klasifikasi diberikan penyimpanan data yang dibagi menjadi set *training data tes*. *Training data set* digunakan dalam membuat model klasifikasi, sementara record data set digunakan untuk menvalidasi model. Model kemudian digunakan untuk mengklasifikasikan dan memprediksi data set baru yang berbeda dari training dan tes data set.

Sistem yang membangun pengklasifikasi adalah salah satu alat yang umum digunakan pada *data mining* [Wu et. al.,2008). Contohnya adalah sistem untuk mengambil kumpulan kasus, dimana masing-masing kasus merupakan salah satu dari beberapa kelas dan nilainya dapat dijelaskan untuk satu set atribut tetap, dan menghasilkan pengklasifikasi yang dapat memprediksi secara akurat kelas tempat kasus baru berada. Atau dapat disimpulkan bahwa teknik klasifikasi merupakan cara untuk memetakan himpunan atribut ke kelas yang sebelumnya sudah didefinisikan.

Teknik Klasifikasi yang dibahas di buku ini antara lain adalah :

1. Decision Tree
2. CART
3. Naïve Bayes
4. Bayes

4.3. Decision Tree

Algoritma *decision tree* paling sering digunakan karena mudah untuk dimengerti dan murah untuk diimplementasikan. *Decision Tree* menyediakan teknik model yang mudah dibandingkan dan disimplfikasikan proses klasifikasinya. Algoritma *decision tree* adalah teknik *data mining* yang secara rekursif membuat partisi data menggunakan pendekatan *depth-first greedy* atau pendekatan *breadth-first* sampai semua data masuk ke dalam kelas tertentu. Struktur *decision tree* dibuat dari root, internal dan node *leaf*. *Entropy* adalah jumlah bit yang dibutuhkan untuk mengekstrak kelas ke dalam sampel data yang dirandom. *Entropi* akan digunakan untuk menentukan root pertama untuk membangun tree. *Entropi* terendah akan dipilih sebagai *root* atau daun dari pohon.

Langkah Klasifikasi

Proses Klasifikasi Data

a. *Learning*: Data *training* dianalisa dengan algoritma klasifikasi. Atribut label kelas adalah keputusan pemberian pinjaman atau tidak, dan model pembelajaran klasifikasi diberikan dalam rule klasifikasi. Jika akurasi dapat diterima maka rule tersebut akan diaplikasikan dalam klasifikasi *decision tree* untuk data baru.

b. Klasifikasi: Data tes digunakan untuk mengestimasi keakuratan dari *rule* klasifikasi. Jika akurasi dapat diterima maka rule dapat diaplikasikan untuk data baru.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut.

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus (Kusrini,2009):

$$Gain(S,A) = Entropi\ (S) - \sum_{i=0}^{n} \frac{|Si|}{|S|} * Entropi\ (Si)$$

Keterangan :

S : himpunan kasus

A : Atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke I

|S| : jumlah kasus dalam S

Perhitungan nilai entropi dapat dilihat pada persamaan (Kusrini,2009)::

$$Entropi\ (S) = \sum_{k=0}^{n} -\text{pi} * \log_2 p_i$$

Keterangan:

S: himpunan kasus

A: fitur

n : jumlah partisi S

p_i : proporsi dari Si terhadap S

4.4. Studi Kasus Decision Tree

Studi Kasus Koperasi Susu di Jawa Barat (Fitriana, 2013)

Sumber Data

Data dikumpulkan dari koperasi susu di Jawa Barat. Koperasi susu menyalurkan kredit sapi bergulir yang didapatkan dari pemerintah dan lembaga swasta. Penelitian ini membutuhkan banyak data, seperti:

1. ID Anggota
2. Nama Peternak
3. TPS
4. TPK
5. Umur
6. Pendapatan
7. Anggota aktif
8. Tidak mempunyai Sisa Kredit atu kredit macet
9. Minimal mempunyai 1 sapi laktasi dan maksimal mempunyai 2 sapi laktasi
10. Menguasai teknik peternakan

Data dikumpulkan dari Januari sampai Februari. Data ini digunakan untuk membuat rule pengambilan keputusan mendapatkan kredit atau tidak dengan menggunakan algoritma *decision tree*. Data bulan Januari digunakan untuk membuat rule formulasi dan data bulan Februari digunakan untuk mengetes rule tersebut.

Pre-Processing Data

Pre-Processing data berfokus pada sumber data untuk *decision tree*. Data harus disiapkan untuk membuat target binary atribut (Ya atau Tidak) dalam penerimaan kredit sapi bergulir. Dan data atribut kualitatif. Klasifikasi data dapat dilihat di bawah ini.

a) Umur

Umur peternak ditransformasi menjadi 3 klasifikasi muda, umur pertengahan dan tua. Klasifikasi umur dibuat agar dapat diketahui data umur peternak yang ingin mengambil kredit.

Tabel 4.1 Klasifikasi umur

Umur	Range (Tahun)
Muda	< 25
Umur Pertengahan	25 - 50
Tua	> 50

b) Pendapatan

Pendapatan peternak ditransformasi menjadi klasifikasi tinggi, sedang dan rendah.

Tabel 4.2 Klasifikasi Pendapatan dalam Rupiah

Pendapatan	Jumlah (Rupiah)
Tinggi	> Rp.4.000.000
Sedang	Rp.2.000.000 – Rp.4.000.000
Rendah	< Rp. 2.000.000

c) Anggota Aktif

Data anggota ditansformasi menjadi klasifikasi anggota aktif dan tidak, berdasarkan data setoran susu peternak dalam 8 bulan terakhir.

Tabel 4.3 Klasifikasi Anggota Aktif

Anggota Aktif	Pengertian
Ya	Dalam 8 bulan terakhir selalu menyetorkan susu setiap hari
Tidak	Dalam 8 bulan terakhir tidak selalu menyetorkan susu

d) Sisa Kredit

Data sisa kredit ditansformasi menjadi klasifikasi ya atau tidak, berdasarkan data masih mempunyai sisa kredit atau tidak.

Tabel 4.4 Klasifikasi Sisa Kredit

Sisa Kredit	Pengertian
Ya	Masih mempunyai sisa kredit
Tidak	Tidak mempunyai sisa kredit

e) Min 1 sapi dan maks 2 sapi

Data jumlah sapi laktasi minimal 1 sapi dan maksimal 2 sapi ditansformasi menjadi klasifikasi ya atau tidak, berdasarkan data masih mempunyai sisa kredit atau tidak.

Tabel 4.5 Klasifikasi Min 1 sapi maks 2 sapi

Min 1 sapi dan maks 2 sapi	Pengertian
Ya	Mempunyai minimal 1 sapi dan maksimal 2 sapi laktasi
Tidak	Tidak mempunyai minimal 1 sapi dan maksimal 2 sapi laktasi

f) Menguasai Teknis Peternakan

Data menguasai teknis peternakan ditansformasi menjadi klasifikasi ya atau tidak, berdasarkan data menguasai teknis peternakan.

Tabel 4.6 Menguasai Teknis Peternakan

Menguasai Teknis Peternakan	Pengertian
Ya	Menguasai teknis peternakan
Tidak	Tidak menguasai teknis peternakan

Pre-Processing data berfokus pada sumber data dari yang digunakan untuk *decision tree*. Data harus disiapkan untuk membuat atribut binary target. (Diterima atau ditolak) dan atribut data kualitatif. Data klasifikasi dapat dilihat di bawah ini.

Tabel 4.7 adalah tabel diagram input output model Perkreditan.

Tabel 4.7 Diagram Input Output Pemodelan Perkreditan

No	Input	Aktivitas	Output
1	Data umur	Mengklasifikasi data umur menjadi muda, usia pertengahan dan tua	Data umur menjadi tua, usia pertengahan

2	Data pendapatan	Mengklasifikasi data pendapatan menjadi tinggi, sedang dan rendah	Data pendapatan menjadi tinggi, sedang dan rendah
3	Data anggota koperasi.	Mengklasifikasi data anggota aktif menjadi ya dan tidak.	Data anggota aktif yang selalu menyetorkan susu selama 8 bulan berturut-turut dan anggota tidak aktif.
4	Data sisa kredit	Mengklasifikasi data sisa kredit menjadi ya dan tidak.	Data sisa kredit menjadi ya dan tidak.
5	Data sapi laktasi min 1 maks 2	Mengklasifikasi data sapi laktasi min 1 maks 2 menjadi ya dan tidak.	Data sapi laktasi min 1 maks 2 menjadi ya dan tidak.
6	Data menguasai teknis peternakan	Mengklasifikasi data menguasai teknis peternakan menjadi ya dan tidak.	Data sapi laktasi min 1 maks 2 menjadi ya dan tidak.

Algoritma *Decision Tree*.

Aturan mengambil keputusan penerimaan kredit sapi bergulir menggunakan Algoritma *Decision Tree*. Bagian penting dari *decision tree* adalah membuat root pertama dari tree. Root pertama diambil dari data atribut. Ada 6 atribut yang akan dihitung yaitu menguasai teknik peternakan, memiliki min 1 maks 2 sapi laktasi, sisa kredit, pendapatan, umur dan anggota aktif. Perhitungan entropi dapat dillihat pada Tabel 4.8 berikut ini.

Tabel 4.8 Perhitungan Node 1

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
1	Total		470	435	35	0.382380675	
	UMUR						
		Tua	137	124	13	0.143835773	0.095872
		Usia Pertengahan	235	223	12	0.290898956	
		Muda	98	88	10	0.475431646	
	PENDAPATAN		470				0.148927
		Tinggi	315	291	24	0.17736852	
		Sedang	126	115	11	0.427398148	
		Rendah	29	29	0	0	
	ANGGOTA AKTIF		470				0.001576448
		Ya	463	428	35	0.38646066	
		Tidak	7	7	0	0	
	Sisa Kredit		470				0.007739237

Node		Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
	Ya	31	25	6	0.708835673	
	Tidak	439	410	29	0.351042301	
Min 1 dan Maks 2						0.13193991
		470				
	Ya	64	31	33	0.999295444	
	Tidak	406	404	2	0.044849614	
Menguasai Teknik Peternakan		470				0.209799887
	Ya	63	30	33	0.998363673	
	Tidak	407	405	2	0.044756902	
		470				

Dari hasil pada Tabel 4.8 dapat diketahui bahwa atribut gain dengan tertinggi adalah menguasai teknik peternakan, yaitu sebesar 0,21. Dengan demikian menguasai teknik peternakan dapat menjadi node akar. Ada dua nilai atribut dari menguasai teknik peternakan yaitu ya dan tidak.Yang tertinggi adalah tidak, sehingga masih perlu dilakukan perhitungan sekali lagi.

Tabel 4.9 Perhitungan Node 2

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
2	Menguasai Teknik Peternakan-Ya		63	30	33	0.998363673	
	UMUR						
		Tua	24	11	13	0.994984828	0.010036
		Usia Pertengahan	26	14	12	0.995727452	
		Muda	13	5	8	0.961236605	
	PENDAPATAN		63				0.332788
		Tinggi	42	20	22	0.998363673	
		Sedang	21	0	0	0	
		Rendah	0	0	0	0	
	ANGGOTA AKTIF		63				0.034808
		Ya	61	28	33	0.995148096	
		Tidak	2	2	0	0	
	Sisa Kredit		63				0.007498
		Ya	14	8	6	0.985228136	
		Tidak	49	22	27	0.992476004	

Node		Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
Min 1 dan Maks 2		63				0.573592
	Ya	40	7	33	0.669015835	
	Tidak	23	23	0	0	
		63				

Dari hasil pada Tabel 4.9 dapat diketahui bahwa atribut gain dengan tertinggi adalah minimal memiliki 1 sapi laktasi dan maksimal memiliki 2 sapi laktasi, yaitu sebesar 0,57. Dengan demikian minimal memiliki 1 sapi laktasi dan maksimal memiliki 2 sapi laktasi dapat menjadi node akar. Ada dua nilai atribut dari minimal memiliki 1 sapi laktasi dan maksimal memiliki 2 sapi laktasi yaitu ya dan tidak. Yang tertingi adalah Ya, sehingga masih perlu dilakukan perhitungan sekali lagi.

Tabel 4.10 Perhitungan Node 3

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
3	Menguasai Teknik Peternakan-Ya dan Min 1 dan Maks 2 sapi		40	7	33	0.669015835	
	UMUR						
		Tua	15	2	13	0.566509507	0.018830
		Usia Pertengahan	16	4	12	0.811278124	
		Muda	9	1	8	0.503258335	
	PENDAPATAN		40				0.062310
		Tinggi	24	2	22	0.41381685	
		Sedang	16	5	11	0.896038233	
		Rendah	0	0	0	0	
	ANGGOTA AKTIF		40				0.065118
		Ya	39	6	33	0.619382195	
		Tidak	1	1	0	0	

Node		Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
Sisa Kredit		40				0.213416
	Ya	12	6	6	1	
	Tidak	28	1	27	0.222284831	

Dari hasil pada Tabel 4.10 dapat diketahui bahwa atribut gain dengan tertinggi adalah sisa kredit, yaitu sebesar 0,213. Dengan demikian sisa kredit dapat menjadi node akar. Ada dua nilai atribut dari sisa kredit yaitu ya dan tidak. Yang tertingi adalah Ya, sehingga masih perlu dilakukan perhitungan sekali lagi.

Tabel 4.11 Perhitungan Node 4

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	Gain
4	Menguasai Teknik Peternakan-Ya dan Min 1 dan Maks 2 - Ya dan Sisa Kredit -Ya		12	6	6	1	
	UMUR		12				
		Tua	4	2	2	1	0.262104
		Usia Pertengahan	5	3	2	0.970950594	
		Muda	3	1	2	0	
	PENDAPATAN		12				**0.654858**
		Tinggi	7	1	6	0.591672779	
		Sedang	5	5	0	0	
		Rendah	0	0	0	0	

Node	Jumlah Kasus (S)	Tidak (S1)	Ya (S2)		Entropi	Gain
ANGGOTA AKTIF	12					0.000000
Ya	12	6	6	1		
Tidak	0	0	0	0		

Dari hasil pada Tabel 4.11 dapat diketahui bahwa atribut gain dengan tertinggi adalah pendapatan, yaitu sebesar 0,65. Dengan demikian pendapatan dapat menjadi node akar. Ada tiga nilai atribut dari pendapatan yaitu timggi, sedang dan rendah. Yang tertingi adalah tinggi, sehingga masih perlu dilakukan perhitungan sekali lagi.

Tabel 4.12 Perhitungan Node 5

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	
5	Menguasai Teknik Peternakan-Ya dan Min 1 dan Maks 2 - Ya dan Sisa Kredit -Ya , Pendapatan –Tinggi		7	1	6	0.591672779	
	UMUR		7				
		Tua	3	1	2	0.918295834	0.198117
		Usia Pertengahan	2	0	2	0	
		Muda	2	1	1	0	
	ANGGOTA AKTIF		7				
		Ya	7	1	6	0.591672779	0.000000
		Tidak	0	0	0	0	

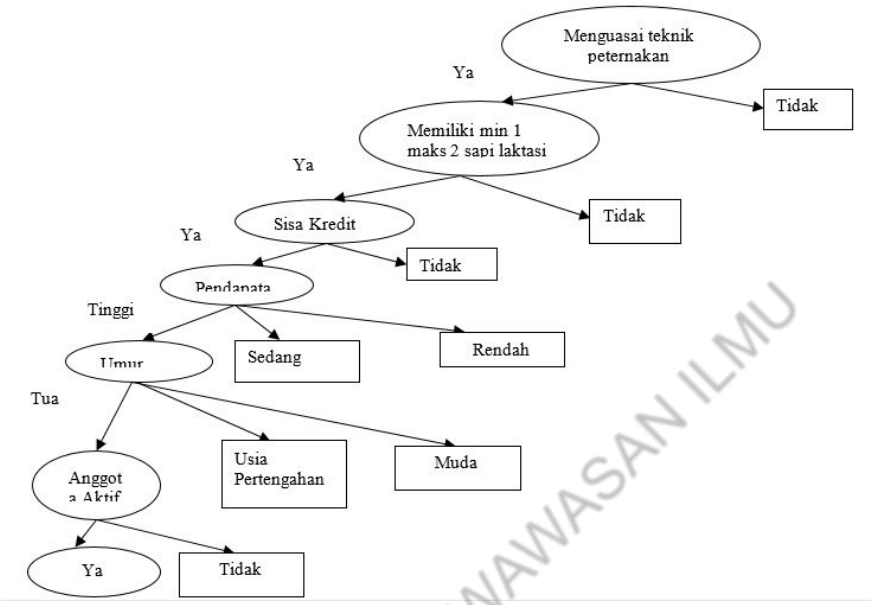
Dari hasil pada Tabel 4.12 dapat diketahui bahwa atribut gain dengan tertinggi adalah umur, yaitu sebesar 0,198. Dengan demikian umur dapat menjadi node akar. Ada tiga nilai atribut dari pendapatan yaitu tua, usia pertengahan dan muda. Yang tertingi adalah tua sehingga masih perlu dilakukan perhitungan sekali lagi.

Tabel 4.13 Perhitungan Node 6

Node			Jumlah Kasus (S)	Tidak (S1)	Ya (S2)	Entropi	
6	Menguasai Teknik Peternakan-Ya, Min 1 dan Maks 2 - Ya, Sisa Kredit -Ya , Pendapatan -Tinggi dan Umur – Tua		3	1	2	0.918295834	
	ANGGOTA AKTIF		3				
		Ya	3	1	2	0.918295834	0.198117
		Tidak	0	0	0	0	

Dari hasil pada tabel dapat diketahui bahwa atribut gain dengan tertinggi adalah anggota aktif yaitu sebesar 0,198. Dengan demikian anggota aktif dapat menjadi node akar. Ada dua nilai atribut dari pendapatan ya dan tida. Yang tertingi adalah ya.

Berikut adalah *decision tree* hasil perhitungan.



Gambar 4.1. Decision Tree

Dari Gambar 4.1 *decision tree, rule* pertama akan didefinisikan, **IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = ya, pendapatan=tinggi, umur = usia pertengahan, anggota aktif = ya THEN penerimaan kredit = ya.** Setelah melakukan *processing data,* ada 11*rule* yang ditemukan dengan penerimaan kredit =ya.

Tabel 4.14 adalah rule penerimaan Rule Penerimaan kredit sapi bergulir dengan target data penerimaan kredit = ya.

Tabel 4.14 Rule Penerimaan kredit sapi bergulir dengan target. data penerimaan kredit = ya

No.	Rules
1	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = ya, pendapatan= tinggi, umur =usia pertengahan, anggota aktif = ya THEN penerimaan kredit = ya.

No.	Rules
2	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan=tinggi, umur = tua, anggota aktif = ya THEN penerimaan kredit = ya.
3	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan=tinggi, umur = muda, anggota aktif = ya THEN penerimaan kredit = ya.
4	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= sedang, umur = usia pertengahan, anggota aktif = ya THEN penerimaan kredit = ya.
5	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= tinggi, umur = usia pertengahan, anggota aktif = ya THEN penerimaan kredit = ya.
6	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= tinggi, umur = tua, anggota aktif = ya THEN penerimaan kredit = ya.
7	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= sedang, umur = muda, anggota aktif = ya THEN penerimaan kredit = ya.
8	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= sedang, umur = tua, anggota aktif = ya THEN penerimaan kredit = ya.
9	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = tidak, pendapatan= sedang, umur = muda, anggota aktif = ya THEN penerimaan kredit = ya.
10	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = ya, pendapatan= tinggi, umur =muda, anggota aktif = ya THEN penerimaan kredit = ya.

No.	Rules
11	IF menguasai teknik peternakan = ya, memiliki min 1 sapi maks 2 sapi laktasi =ya, sisa kredit = ya, pendapatan= tinggi, umur =tua, anggota aktif = ya THEN penerimaan kredit = ya.

Hasil paling penting dari *Decision Tree* adalah rule untuk pengambilan keputusan penerimaan kredit. Hasil pengolahan dengan menggunakan software Weka dapat dilihat pada Gambar 4.2.

```
Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     sapi kredit
Instances:    470
Attributes:   8
              NamaLengkap
              Umur
              Pendapatan
              Anggota Aktif
              Sisa Kredit
              Min 1 dan Maks 2 sapi
              Menguasai Teknis Peternakan
              Mendapat Kredit Sapi
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Menguasai Teknis Peternakan = Ya
|   Min 1 dan Maks 2 sapi  = Ya
|   |   Sisa Kredit = Ya
|   |   |   Pendapatan = Tinggi: Ya (7.0/1.0)
|   |   |   Pendapatan = Sedang: Tidak (5.0)
|   |   |   Pendapatan = Rendah: Tidak (0.0)
|   |   Sisa Kredit = Tidak: Ya (28.0/1.0)
|   Min 1 dan Maks 2 sapi  = Tidak: Tidak (23.0)
```

Menguasai Teknis Peternakan = Tidak: Tidak (407.0/2.0)
Number of Leaves : 6
Size of the tree : 10
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	466	99.1489 %
Incorrectly Classified Instances	4	0.8511 %
Kappa statistic	0.9383	
Mean absolute error	0.0173	
Root mean squared error	0.0952	
Relative absolute error	12.3707 %	
Root relative squared error	36.2403 %	
Total Number of Instances	470	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.995	0.057	0.995	0.995	0.995	0.949	Tidak
	0.943	0.005	0.943	0.943	0.943	0.949	Ya
Weighted Avg.	0.991	0.053	0.991	0.991	0.991	0.949	

=== Confusion Matrix ===

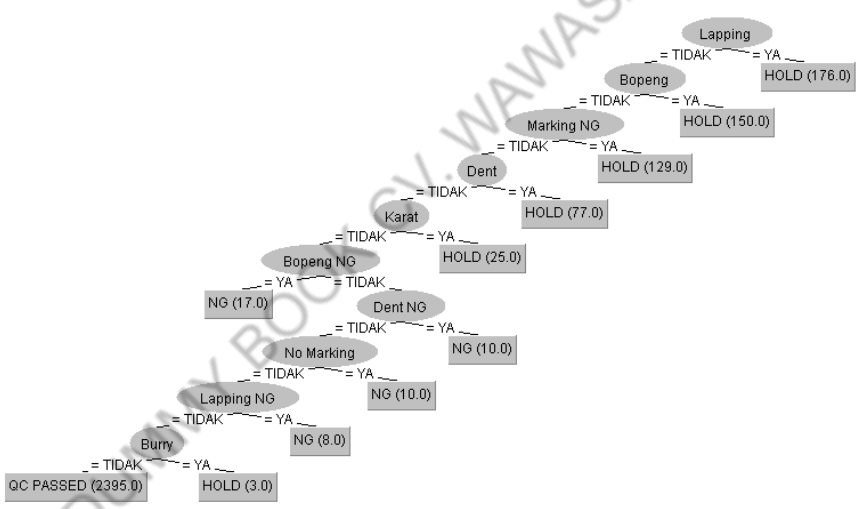
a	b	<-- classified as
433	2	a = Tidak
2	33	b = Ya

Gambar 4.2 Hasil pengolahan menggunakan software Weka

Dari hasil pengolahan *software weka* dapat dilihat *decision tree* dari hasil perhitungan telah benar. Node 1 menguasai teknik peternakan, node 2 min 1 dan maks 2 sapi, node 3 sisa kredit, node 4 pendapatan. Kebenaran rule *decision tree* adalah 99, 15 persen. Dari 470 data, 33 peternak yang layak mendapatkan kredit dan 433 peternak tidak layak mendapatkan kredit. Ada 0,85 klasifikasi yang tidak sesuai dari 470 data ada 4 yang klasifikasinya tidak sesuai.

Studi Kasus Klasifikasi dengan algoritma *decision tree* produk part common rail pada lini forging (Fitriana R. *et. al.*,2020)

Pada tahap ini digunakan data pengamatan pada bulan November 2018. Tahap ini bertujuan untuk menentukan apakah produk sampel tersebut berada di kelas *hold,* NG atau *QC Passed.* Produk dikategorikan *hold* apabila pada produk tersebut ditemukan cacat dan cacat tersebut dapat diperbaiki/*rework*/*repair.* Produk dikaregorikan NG apabila produk tersebut tidak dapat diperbaiki sehingga menjadi *scrap.* Produk dikategorikan *QC Passed* apabila pada produk tersebut tidak ditemukan cacat atau lolos inspeksi. Berdasarkan data pengamatan yang berjumlah 3000 sampel produk, dapat diketahui bahwa sebanyak 45 unit produk dinyatakan NG, 560 unit dinyatakan *hold* dan 2395 unit dinyatakan QC *Passed.* Hasil *decision tree* untuk data pengamatan produk *Common Rail* Tipe 1 dapat dilihat pada Gambar:



Gambar 4.3 Decision Tree Common Rail Type 1

Terdapat sepuluh parameter yang paling berpengaruh pada proses produksi *part Common Rail* Tipe 1. Parameter yang paling berpengaruh yaitu *lapping,* bopeng, *marking* NG, *dent,* karat, bopeng NG, *dent* NG, *no marking, lapping* NG dan *burry.* Terbentuk 11 fungsi yang terbentuk dari *decision tree* pada Gambar. Fungsi yang pertama adalah produk dinyatakan *hold* apabila terdapat cacat *lapping.* Fungsi yang kedua produk dikatakan *hold* apabila tidak terdapat cacat *lapping* namun ditemukan cacat bopeng, dan seterusnya. Tabel *if then rules* dapat dilihat pada Tabel.

Tabel 4.15 If Then Rules

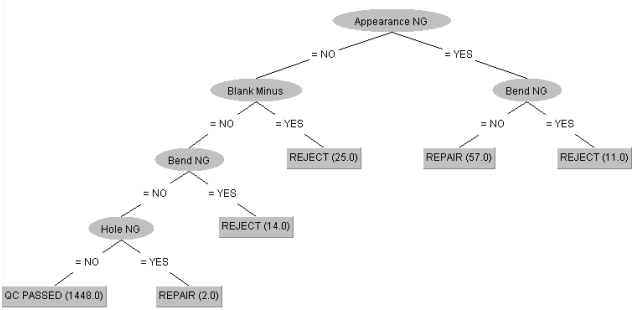
No.	Fungsi	No.	Fungsi	No.	Fungsi
1	If Lapping "YA"	7	IF Lapping "TIDAK"	10	IF Lapping "TIDAK"
	Then Keputusan "HOLD"		Bopeng "TIDAK"		Bopeng "TIDAK"
2	IF Lapping "TIDAK"		Marking NG "TIDAK"		Marking NG "TIDAK"
	Bopeng "YA"		DENT "TIDAK"		DENT "TIDAK"
	Then Keputusan "HOLD"		Karat "TIDAK"		Karat "TIDAK"
3	IF Lapping "TIDAK"	8	Bopeng NG "TIDAK"		Bopeng NG "TIDAK"
	Bopeng "TIDAK"		Dent NG "YA"		Dent NG "TIDAK"
	Marking NG "YA"		Then Keputusan "NG"		No Marking "TIDAK"
	Then Keputusan "HOLD"		IF Lapping "TIDAK"		Lapping NG "TIDAK"
4	IF Lapping "TIDAK"		Bopeng "TIDAK"		Burry "YA"
	Bopeng "TIDAK"		Marking NG "TIDAK"		Then Keputusan "HOLD"
	Marking NG "TIDAK"		DENT "TIDAK"	11	IF Lapping "TIDAK"
	DENT "YA"		Karat "TIDAK"		Bopeng "TIDAK"
	Then Keputusan "HOLD"		Bopeng NG "TIDAK"		Marking NG "TIDAK"
5	IF Lapping "TIDAK"		Dent NG "TIDAK"		DENT "TIDAK"
	Bopeng "TIDAK"		No Marking "YA"		Karat "TIDAK"
	Marking NG "TIDAK"		Then Keputusan "NG"		Bopeng NG "TIDAK"
	DENT "TIDAK"	9	IF Lapping "TIDAK"		Dent NG "TIDAK"
	Karat "YA"		Bopeng "TIDAK"		No Marking "TIDAK"
	Then Keputusan "HOLD"		Marking NG "TIDAK"		Lapping NG "TIDAK"

6	IF Lapping "TIDAK"	DENT "TIDAK"	Burry "TIDAK"
	Bopeng "TIDAK"	Karat "TIDAK"	Then Keputusan "QC PASSED"
	Marking NG "TIDAK"	Bopeng NG "TIDAK"	
	DENT "TIDAK"	Dent NG "TIDAK"	
	Karat "TIDAK"	No Marking "TIDAK"	
	Bopeng NG "YA"	Lapping NG "YA"	
	Then Keputusan "NG"	Then Keputusan "NG"	

Terdapat 10 parameter yang paling berpengaruh pada kualitas produk *Common Rail* Tipe 1 yaitu *lapping,* bopeng, *marking* NG, *dent,* karat, bopeng NG, *dent* NG, *no marking* dan *lapping* NG, *burry*.

Studi Kasus Decision Tree untuk Product Body 2-1 (Saragih J. *et. al.,*2020)

Terdapat empat parameter/atribut yang paling berpengaruh terhadap kualitas *body2-1,* keempat parameter tersebut merupakan kecacatan atribut, yang diantaranya adalah *appearance* NG, *bend* NG, *blank minus,* dan *hole* NG. Empat parameter tersebut menentukan keputusan perusahaan dalam penentuan standarisasi "*QC PASSED*". Berdasarkan data pengamatan yang berjumlah 1557 data atau sampel produk, 1448 produk dinyatakan "*QC PASSED*", 50 produk dinyatakan "*REJECT*" dan 59 produk dinyatakan "*REPAIR*". Standarisasi "*QC PASSED*" dalam perusahaan disajikan dalam *decision tree* pada Gambar 4.4 berdasarkan lima hari pengamatan yang telah dilakukan.



Gambar 4.4 Decision Tree Standarisasi "QC Passed"

Terdapat 6 fungsi yang terbentuk dari *decision tree* pada Gambar 3. Fungsi yang pertama adalah produk dinyatakan *"REJECT"* jika terdapat kecacatan *appearance* NG dan *bend* NG. Fungsi yang kedua adalah produk dinyatakan *"REPAIR"* jika terdapat kecacatan *appearance* NG dan tidak terdapat kecacatan *bend* NG. Fungsi yang ketiga adalah produk dinyatakan "*REJECT"* jika tidak terdapat kecacatan *appearance* NG tetapi terdapat kecacatan *blank minus*. Fungsi yang keempat adalah produk dinyatakan "*REJECT*" jika terdapat kecacatan *appearance* NG, tidak terdapat kecacatan *blank minus* dan terdapat kecacatan *bend* NG. Fungsi yang kelima adalah produk dinyatakan *"REPAIR"* jika tidak terdapat kecacatan *appearance* NG, *blank minus* dan *bend* NG, tetapi terdapat kecacatan *hole* NG. Fungsi yang keenam adalah produk dinyatakan "*QC PASSED"* jika tidak terdapat kecacatan *appearance* NG, *blank minus, bend* NG dan *hole* NG. Tabel *If Then Rules* tertera pada Tabel 5.11.

Tabel 4.16 If Then Rules

No.	Fungsi
1	*If Appearance* NG "YES"
	Bend NG "YES"
	Then Keputusan "*REJECT*"
2	*If Appearance* NG "YES"
	Bend NG "NO"
	Then Keputusan "*REPAIR*"
3	*If Appearance* NG "NO"
	Blank Minus "YES"
	Then Keputusan "*REJECT*"
4	*If Appearance* NG "NO"
	Blank Minus "NO"
	Bend NG "YES"
	Then Keputusan "*REJECT*"
5	*If Appearance* NG "NO"
	Blank Minus "NO"
	Bend NG "NO"
	Hole NG "YES"
	Then Keputusan "*REPAIR*"

No.	Fungsi
6	*If Appearance* NG "NO"
	Blank Minus "NO"
	Bend NG "NO"
	Hole NG "NO"
	Then Keputusan "*QC PASSED*"

Berdasarkan pengolahan *data mining* yang telah dilakukan, karakteristik kualitas yang paling berpengaruh adalah atribut yang diantaranya adalah *appearance* NG (Penampilan produk tidak baik), *balnk minus* (Hasil potongan tidak utuh), *bend* NG (Pembengkokan tidak sempurna) dan *hole* NG (Bentuk lubang tidak baik).

Pembuatan Model *Graphic User Interface* (GUI)

Input yang digunakan dalam pembuatan model GUI adalah fungsi *if then rules* yang dihasilkan dari *decision tree*. Pembuatan dari model GUI ini bertujuan agar pengguna dapat secara langsung mengetahui standarisasi "*QC PASSED*" suatu produk berdasarkan jenis kecacatannya. Model program GUI yang dibuat seperti pada Gambar 4.5.

Gambar 4.5 Model GUI Aktif

4.5. CART (*Classification and Regression Tree*)

Algoritma CART (*Classifcation and Regression Tree*) (Steinberg, 2019) adalah salah satu metode dari *machine learning* (Studer et.al.,

2021) dengan jenis *supervised learning* yaitu penggunaan guru yang ditandai melalui adanya *label* atau kelas (Stanton, 2012). Algoritma CART termasuk kedalam salah satu teknik eksplorasi data yaitu teknik pohon keputusan (Stanton, 2009). CART dikembangkan untuk melakukan analisis klasifkasi pada peubah respon baik yang nominal, ordinal, maupun kontinu. CART menghasilkan suatu pohon klasifikasi jika peubah responnya kategorik dan menghasilkan pohon regresi jika peubah responnya kontinu. Nilai tingkat kesalahan yang paling kecil pada pohon klasifikasi yang dihasilkan akan membuat pohon ini digunakan untuk memperkirakan respon. Prinsip dari metode pohon klasifikasi ini adalah memilah seluruh amatan menjadi dua gugus amatan dan memilah kembali gugus.

4.6. Studi Kasus CART (Alfiandi,2021)

- *Data Transformation*

Data Transformation adalah tahapan untuk dapat mengubah bentuk data yang dilakukan agar set data siap diproses untuk dapat menghasilkan analisis yang lebih baik. Bentuk data kecacatan produk yang diperoleh merupakan data numerik, berupa jumlah produk *reject* pada tiap jenis kecacatan. Agar algoritma CART pada *data mining* berupa *decision tree* dapat diaplikasikan nantinya, maka data perlu dibuat seragam dengan mengubah data numerik menjadi kategorik. Transformasi dilakukan berdasarkan rentang nilai yang telah ditentukan. Tiap atribut jenis kecacatan ditransformasikan menjadi "YES" apabila terdapat kecacatan dan "NO" apabila tidak terdapat kecacatan.

Tahap Modelling

Pembangunan model adalah fase yang perlu dilakukan secara berulang hingga diperoleh model yang tepat dengan parameter yang optimal. Tahap membangun model ini dilakukan dengan menggunakan software MINITAB dan RapidMiner karena pada kedua *software* tersebut tidak dibutuhkan *coding*. Setiap fungsi direpresentasikan dalam bentuk beragam operator fungsional yang dapat dipilih sesuai jenis kajian *data mining* yang diperlukan. Berikut ini terkait langkah-langkah pada tahapan *modelling* yang digunakan dalam penelitian ini :

1. Memilih teknik pemodelan

Data mining dalam penggunaannya memiliki berbagai macam teknik pemodelan yaitu teknik asosiasi, teknik klasterisasi, teknik klasifikasi. Dimana pada setiap model memiliki algoritma untuk menghasilkan sebuah kesimpulan yang didapat melalui data set. Pemilihan algoritma dan teknik yang digunakan juga didasarkan pada fungsi dan tujuan dari penggunaan *data mining*. Pada penelitian ini memiliki tujuan yaitu untuk mengurangi jumlah produk cacat dan pengurangan tingkat *sigma* pada kasus kecacatan produk DB-*Customer Display Product* (DB- CDP).

Dengan jenis data yang digunakan menggunakan kategori "YES" atau "NO" untuk setiap jenis kecacatan yang terjadi. Sehingga data yang ada merupakan data kategori dengan *Respon Binary* yaitu hanya memiliki dua peluang iya atau tidak. Maka dengan tujuan dan jenis data yang digunakan, pada penelitian ini dapat menggunakan dua teknik pemodelan yaitu teknik klasifikasi dan teknik asosiasi, dengan teknik klasifikasi yaitu untuk dapat mengetahui pengelompokkan dari tingkat *reject* yang berdasar pada karakteristik atribut jenis kecacatan. Dan selanjutnya dibuat model asosiasi yang berguna untuk memastikan hasil dari kesesuaian aturan yang didapat, dimana sebelumnya dicari tau terlebih dahulu pola hubungan yang terjadi dan dapat menjadi penyebab dari adanya *reject* yang tinggi.

2. Membangun Model

Tahap selanjutnya yaitu membangun model. Tahapan ini diawali dengan menginput dataset yang sebelumnya sudah dibuat. Pada penilitian ini dataset merupakan data dari *reject record* kecacatan pada produk CDP pada bulan Juni 2021 sampai bulan September 2021.

3. Menjalankan Model

Tahapan selanjutnya apabila model siap dijalankan, maka model dapat di "Run" atau dijalankan. Pada *software* RapidMiner, hasil dari dijalankannya *coding* akan muncul pada kolom *Console*. Pada *software* MINITAB maupun RapidMiner apabila model tidak dapat dijalankan maka akan muncul teks *error* dan memberitau penyebab dari adanya *error*, sehingga dapat dilakukan tinjauan serta dilakukan perbaikan dari *error* yang ditimbulkan.

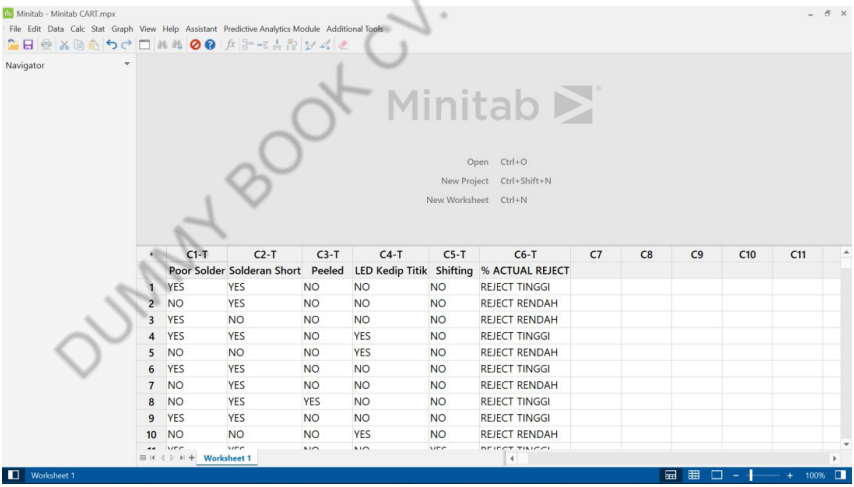
Model Building

- Pembuatan Algoritma CART

Algoritma CART (*Classification and Regression Tree*) merupakan metode teknik klasifikasi data yaitu teknik pohon keputusan. CART dikembangkan untuk dapat menganalisis klasifkasi pada peubah respon baik yang nominal, ordinal, maupun kontinu. Pada penelitian ini mempunyai set data bertipe binomial dengan tipe data kategorik sehingga tidak dapat dilakukan perhitungan untuk regresi, sehingga untuk membuat Decision Tree melalui *Classification* menggunakan *software* MINITAB. Berikut ini langkah untuk membuat *Decision Tree* dari algoritma CART.

1. ***Import Data***

Pada langkah pertama yang dilakukan adalah membuka Software MINITAB, di mana *software* MINITAB yang diperlukan merupakan versi terbaru dengan nama Minitab Statistical Software. Versi ini merupakan versi update 20, dimana fitur untuk pembuatan algoritma CART baru ada pada versi ini. Setelah *Software* terbuka, data dari Microsoft Excel dapat di*copy* kedalam kolom yang terdapat pada MINITAB seperti pada Gambar .

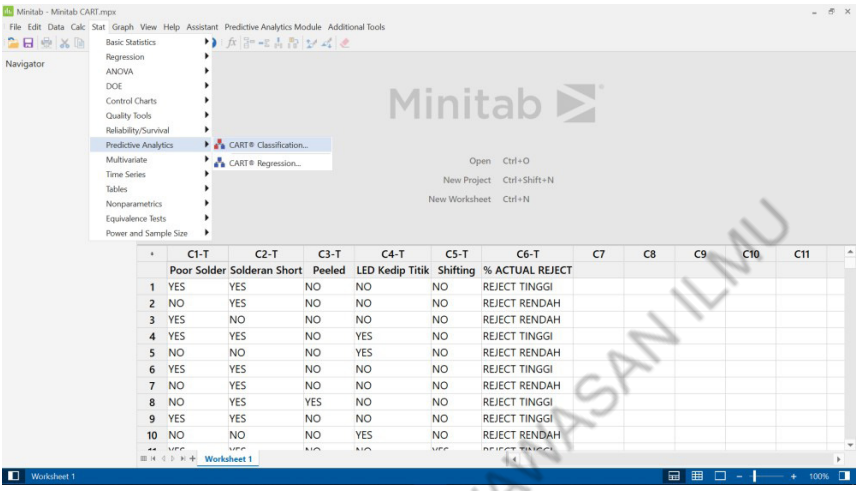


	C1-T	C2-T	C3-T	C4-T	C5-T	C6-T	C7	C8	C9	C10	C11
	Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting	% ACTUAL REJECT					
1	YES	YES	NO	NO	NO	REJECT TINGGI					
2	NO	YES	NO	NO	NO	REJECT RENDAH					
3	YES	NO	NO	NO	NO	REJECT RENDAH					
4	YES	YES	NO	YES	NO	REJECT TINGGI					
5	NO	NO	NO	YES	NO	REJECT RENDAH					
6	YES	YES	NO	NO	NO	REJECT TINGGI					
7	NO	YES	NO	NO	NO	REJECT RENDAH					
8	NO	YES	YES	NO	NO	REJECT TINGGI					
9	YES	YES	NO	NO	NO	REJECT TINGGI					
10	NO	NO	NO	YES	NO	REJECT RENDAH					

Gambar 4.6 Tampilan Awal *Software Minitab* setelah Dimasukkannya *Dataset*

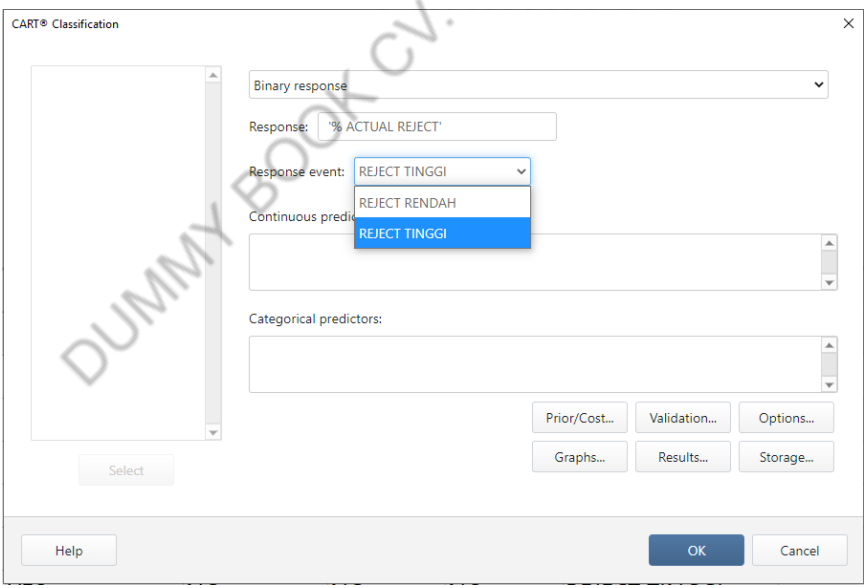
Tahap selanjutnya yaitu pemilihan algoritma CART *classification*, dengan langkah yaitu klik Stat, selanjutnya klik Predictive Analysis, dan kemudian Klik CART *Classification*. Hal tersebut dikarenakan

pada penilitian ini memiliki jenis data kategorik yaitu “YES” atau “NO”, sedangkan untuk CART Regresion digunakan untuk jenis variabel respon berupa data numerikal kontinu. Berikut ini Gambar untuk gambaran mengenai pemilihan algoritma CART.



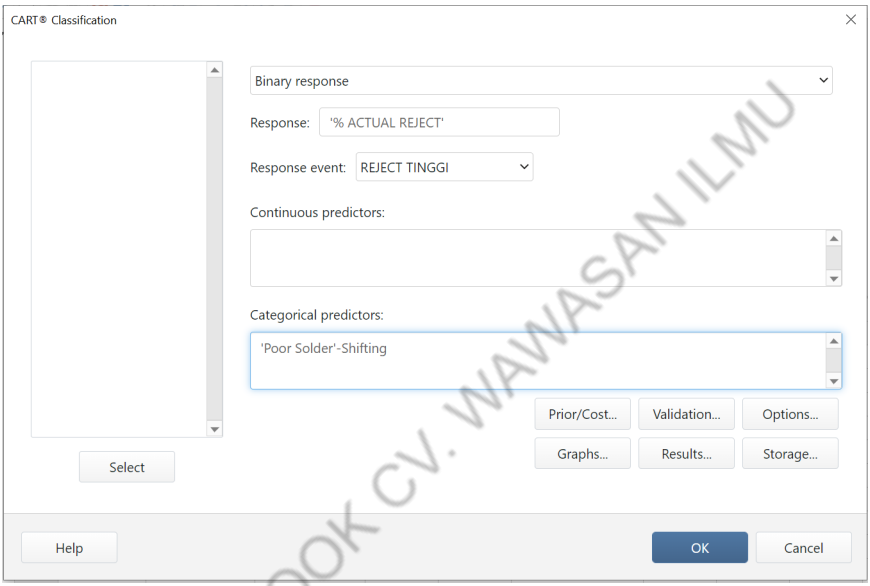
	C1-T	C2-T	C3-T	C4-T	C5-T	C6-T	C7	C8	C9	C10	C11
	Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting	% ACTUAL REJECT					
1	YES	YES	NO	NO	NO	REJECT TINGGI					
2	NO	YES	NO	NO	NO	REJECT RENDAH					
3	YES	NO	NO	NO	NO	REJECT RENDAH					
4	YES	YES	NO	YES	NO	REJECT TINGGI					
5	NO	NO	NO	YES	NO	REJECT RENDAH					
6	YES	YES	NO	NO	NO	REJECT TINGGI					
7	NO	YES	NO	NO	NO	REJECT RENDAH					
8	NO	YES	YES	NO	NO	REJECT TINGGI					
9	YES	YES	NO	NO	NO	REJECT TINGGI					
10	NO	NO	NO	YES	NO	REJECT RENDAH					

Gambar 4.7 Pemilihan Algoritma CART *Classification*



Gambar 4.8 Pemilihan Respon pada Algoritma CART *Classification*

Selanjutnya terdapat pemilihan respon, di mana respon yang dipilih bertipe Binary, di mana hanya memiliki jenis tipe respon yaitu "YES" dan "NO" untuk 5 jenis kecacatan. Dan pada response dipilih "% ACTUAL *REJECT*" dengan response event yang dipilih adalah *REJECT* TINGGI, yang berarti pada penggunaan algoritma CART ini akan mencari penyebab dari adanya respon *Reject* Tinggi dari set data kecacatan.



Gambar 4.9 Pemilihan *Categorical Predictor* Algoritma CART *Classification*

Selanjutnya pada *categorical predictors* diisi jenis kecacatan yang ada, dimana terdapat 5 jenis kecacatan yaitu solderan *Short,* LED kedip titik, *Poor solder*, *peeled*, dan *shifting*, kemudian klik OK.

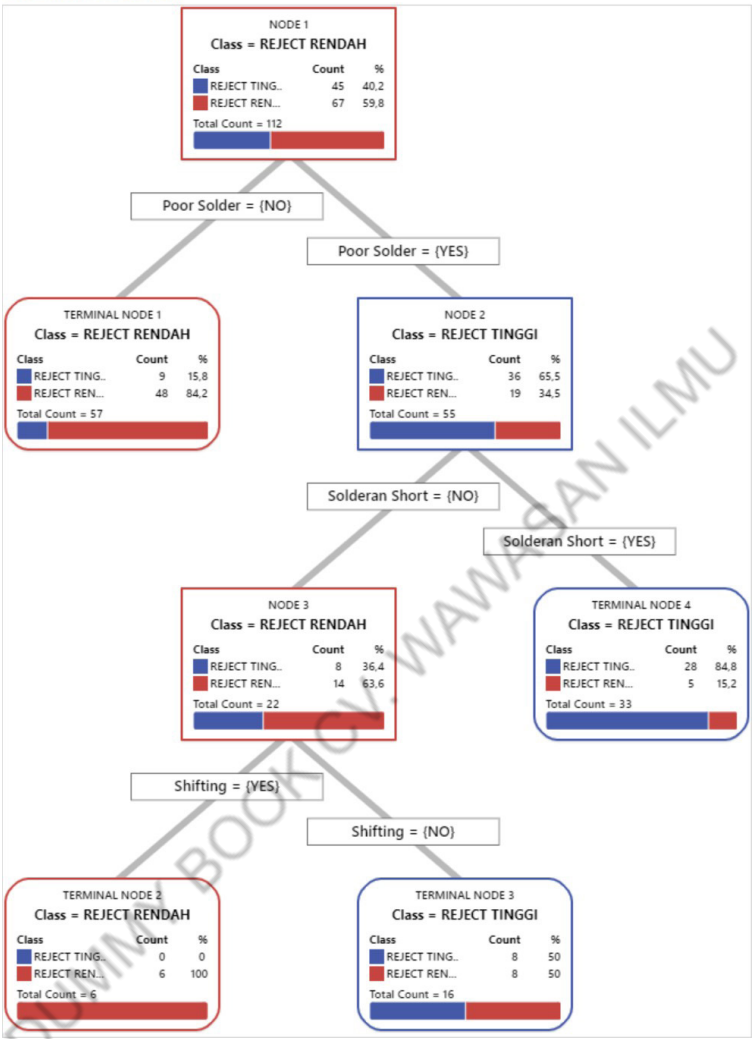
- *Output* Teknik Klasifikasi

Klasifikasi merupakan penyusunan bersistem pada suatu kelompok berdasarkan kaidah atau standar yang ditetapkan. Dimana pada penelitian ini teknik klasifikasi menggunakan algoritma CART (*Classification and Regression Tree*). Metode ini adalah metode statistika nonparametrik, yaitu sistem asumsi tidak melibatkan asumsi dari nilai populasi untuk menggambarkan

hubungan variabel respon (dependen) dengan variabel prediktor (variabel independen), sehingga tidak memiliki asumsi distribusi variabel independen yang perlu dipenuhi. Adapun tujuan utama penggunaan algoritma CART yaitu memudahkan untuk melakukan eksplorasi terhadap set data *Reject Record* dan untuk dapat mengambil keputusan pada struktur data yang dilihat secara visual dengan bentuk *decision tree*.

Decision tree merupakan grafik pohon keputusan untuk mengklasifikasikan jenis atribut dari kumpulan data. Penggunaan dari grafik pohon juga dapat memudahkan pemahaman bagi pembacanya. Teknik ini akan menghasilkan cabang biner, yaitu pada pemilahan data akan dilakukan pemilihan data yang terkumpul dalam suatu ruang yang disebut simpul atau node. Node tersebut akan membuat dua anak yang disebut *child nodes*. Pada prosedur ini dilakukan secara *recursive*, di mana pemisahan dilakukan berulang hingga memenuhi kriteria tertentu. Pada metode ini juga dilakukannya *partitioning*, yaitu proses klasifikasi dilakukan dengan memilah beberapa data menjadi beberapa bagian. Pada titik paling atas *Decision tree* merupakan akar atau disebut dengan *root node*. Dan titik paling akhir yang tidak memiliki cabang lagi disebut dengan daun (*leaf node*) atau disebut juga dengan terminal node.

Pemisahan node atau *node splitting* untuk membuat *child nodes* pada CART menggunakan indeks gini. Indeks Gini digunakan untuk dapat menganalisis skenario secara *real-time*, dengan data asli yang diambil dari analisis *real-time*, dengan cara menghitung jumlah probabilitas spesifikasi tertentu yang diklasifikasikan pada saat pemisahan biner, hingga menghasilkan Terminal Node.



Gambar 4.10 Tampilan *Decision Tree* Algoritma CART

Pada Gambar = dihasilkan *optimal tree diagram* dengan algoritma CART menggunakan teknik klasifikasi dari set data *reject record* produk CDP pada bulan Juni hingga bulan September 2021 dengan hasil *output* berupa klasifikasi *reject*.

Hasil keputusan tersebut bergantung banyaknya atribut yang ada yaitu "YES" atau "NO". Berdasarkan pada hasil *decision tree*, di dapat dua jenis atribut kecacatan yang paling berpengaruh atas jenis tingginya persentase jumlah cacat produk CDP. Dua jenis atribut kecacatan yang berpengaruh adalah *Poor Solder* (solderan kurang) dan *solderan Short.* Pada klasifikasi ini dibentuk cabang berdasarkan nilai persentase tertinggi. Seperti pada gambar 4.48, *root node* merupakan atribut respon yang dipilih yaitu %*Actual Reject*. Atribut tersebut merepresentasikan informasi yang potensial untuk pemecahan cabang selanjutnya.

Diagram tersebut menghasilkan 4 Node, di mana Node 1 merupakan *class object* "*REJECT* RENDAH", yang didapat melalui perbandingan pada *%Actual Reject*, yaitu antara banyaknya persentase *class reject* rendah yaitu 59,8% dengan persentase *class reject* tinggi yaitu 40,2%. Karena persentase *reject* rendah lebih besar, maka node 1 masuk kedalam class *reject* rendah. Pada Node 2 dihasilkan *class Reject* tinggi. Node ini didapat dari *if-then* rules, jika adanya kecacatan *Poor solder* (*"POOR SOLDER* = {*YES*}") maka keputusan yang didapat yaitu "*REJECT* TINGGI" sebesar 65,5%. Selanjutnya pada Node 3 dengan aturan *If "POOR SOLDER* = {*YES*}; *SOLDERAN SHORT* = {*NO*}" menghasilkan keputusan "*REJECT* RENDAH". Node 4 memiliki aturan *if* "*POOR SOLDER* = {*YES*}; SOLDERAN *SHORT* = {*YES*}", yang berarti jika suatu modul memiliki kecacatan *Poor* solder dan solderan *Short*, maka keputusan yang di dapat yaitu "*REJECT* TINGGI". Pada terminal Node 3 memiliki class "*REJECT* TINGGI", dengan aturan jika mengalami *Poor solder*, dan tidak terdapat solderan *Short* maupun shifting. Untuk memudahkan untuk membaca pohon keputusan tersebut maka dibuat Tabel 4.8 sebagai hasil algoritma CART. Dengan kesimpulan *rules* yaitu *Poor* Solder, dan Solderan *Short* merupakan atribut kecacatan yang berpotensi untuk menyebabkan *Reject* yang tinggi.

Tabel 4.17 *If Then Rules* hasil Algoritma CART

No	Jika (*If*)	Maka (*Then*)
1	Poor Solder = {YES}	REJECT TINGGI
2	Poor Solder = {NO}	REJECT RENDAH
3	Poor Solder = {YES}; Solderan Short = {NO}	REJECT RENDAH
4	Poor Solder = {YES}; Solderan Short = {YES}	REJECT TINGGI
5	Poor Solder = {YES}; Solderan Short = {NO}; Shifting = {YES}	REJECT RENDAH
6	Poor Solder = {YES}; Solderan Short = {NO}; Shifting = {NO}	REJECT TINGGI

Berdasarkan pemodelan yang dilakukan dengan menggunakan *Decision Tree* dan *FP-Growth* dapat disimpulkan bahwa dengan adanya 2 atribut dominan yaitu *Poor* solder dan Solderan *Short* pada *Decision Tree*, dan kedua atribut tersebut terdapat pada aturan asosiasi *FP-Growth*. Sehingga kedua atribut kecacatan tersebut yaitu *Poor* solder dan Solderan *Short* pada *Decision Tree* berpotensi mengakibatkan *Reject* yang tinggi, sehingga analisis pada permasalahan difokuskan pada kedua jenis kecacatan tersebut.

Tahap Evaluation

Tahapan selanjutnya yaitu evaluasi dari hasil model algoritma *data mining* yang digunakan sesuai kriteria kesuksesan yang sudah ditentukan pada tahapan sebelumnya. Evaluasi ini dapat dihitung melalui performansi yang didapat maupun tingkat akurasi. Apabila hasil evaluasi kurang baik, maka perlu dilakukannya suatu perubahan yang mengakibatkan, untuk memungkinnya kembali ke fase sebelumnya atau *re-modelling* sehingga akan mendapatkan hasil lebih maksimal.

Evaluasi Model dengan Tabel *Confusion Matrix*

Pada hasil *output* yang dihasilkan oleh *Software* MINITAB pada pembentukan *Decision Tree* akan menghasilkan Tabel *Confusion Matrix*. Tabel ini akan menghasilkan nilai data *Training* dan data *Testing*. Data *training* adalah data yang dilatih untuk membuat model, sedangkan data *testing* merupakan data yang tidak terlibat untuk membuat model. Apabila dihasilkan data training semakin besar (mendekati 100%) maka semakin baik kinerja model yang dijalankan. Berikut ini Tabel *Confusion Matrix* yang dihasilkan.

Tabel 4.18 *Confusion Matrix*

		Predicted Class (Training)			Predicted Class (Test)		
		REJECT	REJECT	% Correct	REJECT	REJECT	% Correct
Actual Class	Count	TINGGI	RENDAH		TINGGI	RENDAH	
REJECT TINGGI (Event)	45	36	9	80,0	36	9	80,0
REJECT RENDAH	67	13	54	80,6	13	54	80,6
All	112	49	63	80,4	49	63	80,4

Statistics	Training (%)	Test (%)
True positive rate (sensitivity or power)	80,0	80,0
False positive rate (type I error)	19,4	19,4
False negative rate (type II error)	20,0	20,0
True negative rate (specificity)	80,6	80,6

Pada tabel *confusion matrix* didapatkan 4 nilai statistik yaitu *True Positive Rate* (TP), *False Positive Rate* (FP), *False Negative Rate* (FN) dan *True Negative Rate* (TN). Pada tabel tersebut diasumsikan bahwa nilai "*REJECT* TINGGI" diasumsikan positif, dan "*REJECT* RENDAH" diasumsikan negatif.

True Positive Rate (TP) di mana *reject* tinggi diprediksi dengan benar. *False Positive Rate* (FP) yaitu didapat kesalahan prediksi *reject* rendah yang sebenarnya *reject* tinggi, FP merupakan *error* tipe 1. *False Negative Rate* (FN) yaitu didapat kesalahan prediksi *reject* tinggi yang sebenarnya *reject* rendah, FN merupakan *error* tipe 2. Dan *True Negative Rate* (TN) dimana *reject* rendah diprediksi dengan benar. Sehingga TP dan TN merupakan hasil yang sesuai dengan prediksinya. Nilai dari 4 objek tersebut dapat dijadikan acuan untuk matrix evaluasi yang meliputi perhitungan *accuracy*, *error rate*, *sensitivity*, *specificity*, *precission* dan *recall*. Berikut ini perhitungannya:

$$\text{Accurancy} = \frac{TP + TN}{P + N} = \frac{80 + 80{,}6}{80 + 80{,}6 + 19{,}4 + 20} = 80{,}3\%$$

$$\text{Error rate} = \frac{FP + FN}{P + N} = \frac{19{,}4 + 20}{80 + 80{,}6 + 19{,}4 + 20} = 19{,}7\%$$

$$\text{Sensitivity} = \frac{TP}{P} = \frac{80}{80 + 20} = 80\%$$

$$\text{Spesificity} = \frac{TN}{N} = \frac{80{,}6}{80{,}6 + 19{,}4} = 80{,}6\%$$

$$\text{Reecal} = \frac{TP}{TP + FN} = \frac{80}{80 + 20} = 80\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{80}{80 + 19{,}4} = 0{,}805 = 80{,}5\%$$

Berdasarkan perhitungan tersebut didapatkan nilai *accuracy*, *error rate*, *sensitivity*, *specificity*, *precision* dan *recall*. Nilai akurasi merupakan ukuran yang spesifikasi yang digunakan untuk standar umum ukuran dari model klasifikasi yang digunakan, dalam penelitian ini yaitu *Decision tree.* Pada perhitungan didapat nilai akurasi sebesar 80,3%, dan menunjukan bahwa model memiliki akurasi yang tinggi dengan nilai *error rate* sebesar 19,7%. Nilai *error rate* merupakan persentase kesalahan yang dapat terjadi pada model yang digunakan, nilai ini juga dapat dihitung melalui 1 dikurang dengan nilai akurasi. Selanjutnya didapat nilai *Sensitivity* sebesar 80%, nilai ini merupakan kemampuan memprediksi benar terhadap *Reject* yang tinggi, sehingga walaupun ketepatan akurasi yang didapat 80,3%, tetapi didapat nilai prediksi benar terhadap *reject* tinggi yang memiliki persentase yang sedikit lebih kecil. *Precision* yaitu nilai persentase akurasi antara data yang diminta positif yaitu *reject tinggi* dengan hasil prediksi yang didapat oleh model, dan didapat nilai persentase presisi sebesar 80,5%. *Specificity* dan *recall* adalah *true rate* yaitu kemampuan untuk memprediksi dengan benar pada *reject* yang rendah maupun *reject* yang tinggi, dan didapatkan nilai 80% untuk *reject* tinggi dan 80,6% untuk *reject* rendah, sehingga data *testing* dianggap dapat diprediksi dengan benar.

Pada tahap objektif bisnis, memiliki tujuan untuk mengurangi jumlah produk CDP (*Character Display Product*) cacat, sehingga memerlukan analisis terhadap hasil yang telah didapatkan. Berdasarkan tahap *modelling* baru didapatkan atribut kecacatan yang mampu mengakibatkan adanya *reject* yang tinggi, tetapi belum menjelaskan penyebab dari masalah yang ada.

4.7. Klasifikasi dengan R Studio

Berikut adalah penjelasan Klasifikasi dengan R Studio.

R Markdown install packages RMarkdown di console # Cara membuat chunks. Anda dapat dengan cepat menambahkan chunks seperti ini ke dalam file anda dengan keyboard shortcut *Ctrl + Alt + I* (OS X: Cmd + Option + I) *Add Chunk* command dalam editor toolbar atau dengan mengetikkan chunk delimiters *{r} dan*

summary(iris)

```
##  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
## Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
## 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
## Median :5.800   Median :3.000   Median :4.350   Median :1.300
## Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
## 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
## Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
## Species
##  setosa :50
##  versicolor :50
## virginica :50
##
####
```

head(iris)

```
## 1   5.1   3.5   1.4   0.2   setosa   l Width Species
## 2   4.9   3.0   1.4   0.2   setosa
## 3   4.7   3.2   1.3   0.2   setosa
## 4   4.6   3.1   1.5   0.2   setosa
## 5   5.0   3.6   1.4   0.2   setosa
## 6   5.4   3.9   1.7   0.4   setosa
```

Gambar 5. 1 Mencoba RWeka packages untuk classification

Provided the Weka classification tree learner implements the "Drawable" interface (i.e., provides a graph method), write_ to_dot can be used to create a DOT representation of the tree for visualization via Graphviz or the Rgraphviz package.

J48 generates unpruned or pruned C4.5 decision trees (Quinlan, 1993)

LMT implements "Logistic Model Trees" (Landwehr, 2003; Landwehr et al., 2005)

M5P (where the P stands for 'prime') generates M5 model trees using the M5' algorithm, which was introduced in Wang & Witten (1997) and enhances the original M5 algorithm by Quinlan (1992) DecisionStump implements decision stumps (trees with a single split only), which are frequently used as base learners for meta learners such as Boosting

install packages RWeka di console

install.packages(RWeka)

```
library(RWeka) # load data data <- iris head(data)
##  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Membuat Model Klasifikasi dengan salah satu algorithm di RWeka Packages

```
# fit model
fit <- J48(Species~., data=iris)# summarize the fit summary(fit)

##
## === Summary ===##
## Correctly Classified Instances
                                147               98     %
## Incorrectly Classified Instances    3          2      %
## Kappa statistic                     0.97
```

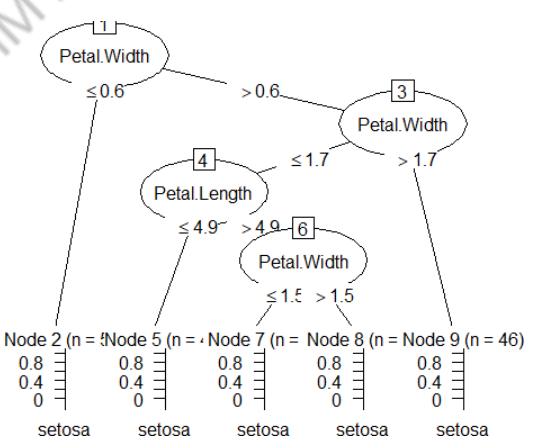
```
## Mean absolute error                0.0233
## Root mean squared error            0.108
## Relative absolute error            5.2482 %
## Root relative squared error        22.9089 %
## Total Number of Instances##        150
## === Confusion Matrix ===##
##  a  b  c   <-- classified as
```

Gambar 5. 2 Menampilkan hasil prediksi klasifikasi dengan tabel

```
table(iris$Species, predict(fit))
##
##              setosa versicolor virginica
##  setosa          50          0         0
##  versicolor       0         49         1
##  virginica        0          2        48
```

Gambar 5. 3 Menggunakan packages partykit untuk menggambarkan hasil klasifikasi install packages partykit di console

```
## visualization
## use partykit package
if(require("partykit", quietly = TRUE))
plot(fit)
```



Gambar 4.11 Membagi ke dalam data training dan data testing Cross validation

Kita akan melakukan cross validation (membagi data menjadi bagian-bagian tertentu). Pada kesempatan kali ini, kita membagi dtaa menjadi 2, yakni data train dan test dengan proporsi data train adalah 80% dan data test 20%.

```
set.seed(100) #Set the seed of R's random number generator, which is useful for creating simulations orrandom objects that can be reproduced
RNGkind(sample.kind = "Rejection") #sample.kind can be "Rounding" or "Rejection", or partial matchesto these. The former was the default in versions prior to 3.6.0: it made sample noticeably non-uniform onlarge populations, and should only be used for reproduction of old results.
idx_iris <- sample(nrow(iris), nrow(iris)*0.8) #membagi 80% dan 20%
train_iris <- iris[idx_iris,]
test_iris <- iris[-idx_iris,]
```

Gambar 4.12 Melihat data training

```
head(train_iris)
##     Sepal.Length Sepal.Width Petal.Length Petal.Width     Species
## 102          5.8         2.7          5.1         1.9   virginica
## 112          6.4         2.7          5.3         1.9   virginica
## 4            4.6         3.1          1.5         0.2      setosa
## 55           6.5         2.8          4.6         1.5  versicolor
## 70           5.6         2.5          3.9         1.1  versicolor
## 98           6.2         2.9          4.3         1.3  versicolor
```

```
head(test_iris)
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 10          4.9         3.1          1.5         0.1  setosa
## 11          5.4         3.7          1.5         0.2  setosa
## 13          4.8         3.0          1.4         0.1  setosa
## 21          5.4         3.4          1.7         0.2  setosa
```

Gambar 4.13 Melihat data testing

Decision tree di R

Model decision tree yang akan kita buat kali ini merupakan model dari library party kit. Adapun fungsi yang bisa digunakan adalah *ctree()*. Sebelumnya, kita load terlebih dahulu library yang digunakan.

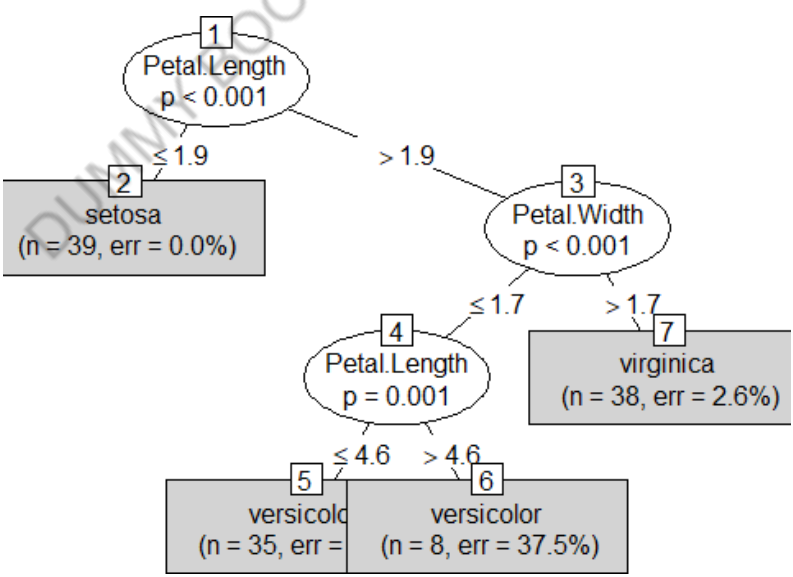
Info tentang CTREE(Conditional Inference Trees: Recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables in a conditional inference framework. https://www.rdocumentation.org/packages/partykit/versions/1.2-15/topics/ctree

```
#install packages caret di console
library(caret) # mengevaluasi model dengan confusion matrix
## Loading required package: ggplot2
## Loading required package: lattice
library(partykit) # pemodelan decision tree
#PEMBUATAN MODEL
model_iris <- ctree(formula = Species ~ ., data = train_iris)
```

Gambar 4.14 Visualisasi model

Untuk membuat visualisasi dari model yang sudah dibuat, kita bisa menggunakan fungsi plot().

```
plot(model_iris, type = "simple")
```



Gambar 4.15 Hasil Model

Dari hasil pemodelan kita bisa mengetahui jika petal length ≤ 1.9 maka termasuk spesies sentosa. Ada 39 observasi yang mengikuti aturan tersebut. Jika petal length > 1.9 maka dilihat lagi petal width. Jika petal width > 1.7 maka termasuk spesies virginica dan begitu seterusnya.

Selanjutnya model yang telah dibuat digunakan untuk memprediksi data baru:

```
pred_iris <- predict(model_iris, test_iris)summary(pred_iris)
##
## Setosa   versicolor   virginica
      11          11          8
class(pred_iris)
## [1] "factor"
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':##
## filter, lag
## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
pred_result <- cbind(test_iris, pred_iris)head(pred_result)
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species pred_iris
## 1          5.1         3.5          1.4         0.2  setosa     setosa
## 6          5.4         3.9          1.7         0.4  setosa     setosa
## 10         4.9         3.1          1.5         0.1  setosa     setosa
## 11         5.4         3.7          1.5         0.2  setosa     setosa
## 13         4.8         3.0          1.4         0.1  setosa     setosa
## 21         5.4         3.4          1.7         0.2  setosa     setosa
```

Gambar 4.16 Prediksi model

Semua spesies hasil prediksi benar dengan spesies data aktualnya.

Evaluasi model

Pada kasus klasifikasi, banyak evaluasi yang bisa digunakan. Pada kasus kali ini, kamu akan menggunakan sebuah *confusion matrix* untuk melihat seberapa baik model dalam memprediksi data yang belum ia lihat. Terdapat beberapa metric yang bisa diperhatikan dalam confusion matrix, pada kesempatan kali ini, kita akan membahas mengenai metric akurasi.

```
confusionMatrix(pred_iris, test_iris$Species)
## Confusion Matrix and Statistics
##
##          Reference
## Prediction setosa versicolor virginica
##  setosa         11         0         0
##  versicolor      0        10         1
##  virginica       0         0         8
##
## Overall Statistics ##
##          Accuracy : 0.9667
##          95% CI : (0.8278, 0.9992)
##          No Information Rate : 0.3667
##          P-Value [Acc > NIR] : 4.476e-12 ##
##          Kappa : 0.9497 ##
## Mcnemar's Test P-Value : NA ##
## Statistics by Class:
##
##          Class: setosa Class: versicolor Class: virginica
## Sensitivity    1.0000            1.0000           0.8889
## Specificity    1.0000            0.9500           1.0000
## Pos Pred Value         1.0000    0.9091           1.0000
## Neg Pred Value         1.0000    1.0000           0.9545
## Prevalence     0.3667            0.3333           0.3000
## Detection Rate         0.3667    0.3333           0.2667
## Detection Prevalence   0.3667    0.3667           0.2667
## Balanced Accuracy      1.0000    0.9750           0.9444
```

Gambar 4.17 Evaluasi Model

Dari hasil confusion matrix di atas, didapat akurasi 0.9667 atau model kita memiliki akurasi 96.67% terhadap data test. Selanjutnya juga diperlukan untuk mengecek model yang dibuat dengan menggunakan data train. Hal ini digunakan untuk mengetahui kondisi apakah model mengalami *under-fitting* atau *over-fitting*. *Overfitting* mempelajari model terlalu baik namun menjadi masalah karena tujuan kita adalah ingin mendapat tren dari sebuah dataset. Model ini menangkap semua tren tetapi bukan tren yang dominan. Model pun tidak bisa menghasilkan output yang reliable karena tidak memiliki kemampuan untuk dapat memprediksi kemungkinan output untuk input yang belum pernah diketahui.

Underfitting tidak mempelajari data dengan baik. Underfitting merupakan keadaan dimana model machine learning tidak bisa mempelajari hubungan antara variabel dalam data serta memprediksi atau mengklasifikasikan data point baru. https://algorit.ma/blog/data-science/overfitting- underfitting/#:~:text=Apa%20Itu%20Underfitting%3F,atau%20mengklasifikasikan%20data%20point%20baru.

```
pred_iris_train <- predict(model_iris, train_iris) pred_iris_train
##    102        112        4          55         70          98         135
## virginica virginica  setosa  versicolor versicolor versicolor versicolor
##      7         43        140        51         25         2          68
## setosa      setosa  virginica  versicolor    setosa     setosa   versicolor
##    137         48        32         85         91         121        16
## virginica    setosa   setosa   versicolor    versicolor    virginica setosa
##    116         66        146        93         45         30        124
## virginica  versicolor   virginica    versicolor    setosa    setosa virginica
##    126         87        95         97         120        29        92
## virginica  versicolor  versicolor versicolor versicolor  setosa versicolor
##   31          54        41         105        113        24        142
## setosa   versicolor   setosa   virginica   virginica   setosa   virginica
##    143         63        65         9          150        20        14
## virginica  versicolor  versicolor  setosa   virginica   setosa    setosa
##      78         88        3        36         27         46         59
## versicolor   versicolor   setosa   setosa   setosa   setosa   versicolor
##      96         69        47         147        129        136        12
## versicolor   versicolor   setosa  virginica virginica  virginica  setosa
```

```
##    141        130         56        22        82         53         99
## virginica versicolor versicolor setosa versicolor versicolor versicolor
##      5         44         28         52        139         42         15
## setosa  setosa  setosa versicolor virginica  setosa  setosa
##     57         75         37         26        110        100        149
## versicolor versicolor setosa  setosa  virginica versicolor virginica
##    132        107         35         58        127        111        144
## virginica versicolor setosa  versicolor virginica virginica virginica
##     86        114         71        123        119         18          8
## versicolor virginica virginica virginica virginica setosa  setosa
##    128         83        138         19        115         23         89
## virginica versicolor virginica setosa  virginica setosa  versicolor
##     62         80        104         40         17         94        133
## versicolor versicolor virginica setosa  setosa  versicolor virginica
##     60         81        118        125        122         49        148
## versicolor versicolor virginica virginica virginica setosa  virginica
##     61
## versicolor
## Levels: setosa versicolor virginica
confusionMatrix(pred_iris_train, train_iris$Species) ## Confusion Matrix and Statistics
##
## Reference
## Prediction setosa versicolor virginica
 ## setosa         39         0         0
 ## versicolor      0         39        4
 ## virginica       0         1         37
 ##
```

```
## Overall Statistics##
##        Accuracy : 0.9583
##        95% CI : (0.9054, 0.9863)
##        No Information Rate : 0.3417
##        P-Value [Acc > NIR] : < 2.2e-16##
##        Kappa : 0.9375##
## Mcnemar's Test P-Value : NA##
## Statistics by Class:
##
##        Class: setosa Class: versicolor Class: virginica


## Sensitivity          1.000    0.9750          0.9024
## Specificity          1.000    0.9500          0.9873
## Pos Pred Value       1.000    0.9070          0.9737
## Neg Pred Value       1.000    0.9870          0.9512
## Prevalence           0.325    0.3333          0.3417

## Detection Rate         0.325      0.3250           0.3083
## Detection Prevalence   0.325      0.3583           0.3167
## Balanced Accuracy      1.000      0.9625           0.9449
?confusionMatrix
```

```
## starting httpd help server ... donetrain_iris$Species
##  [1] virginica  virginica  setosa   versicolor versicolor versicolor
##  [7] virginica  setosa setosa    virginica  versicolor setosa
## [13] setosa   versicolor virginica  setosa   setosa   versicolor
## [19] versicolor virginica  setosa   virginica versicolor virginica
## [25] versicolor setosa   setosa virginica virginica versicolor
## [31] versicolor versicolor virginica  setosa   versicolor setosa
## [37] versicolor setosa   virginica  virginica  setosa   virginica
## [43] virginica  versicolor versicolor setosa   virginica   setosa
## [49] setosa   versicolor versicolor setosa   setosa   setosa
## [55] setosa   versicolor versicolor versicolor setosa   virginica
## [61] virginica  virginica  setosa  virginica  virginica  versicolor
## [67] setosa  versicolor versicolor versicolor setosa  setosa
## [73] setosa   versicolor virginica  setosa   setosa   versicolor
## [79] versicolor setosa   setosa   virginica versicolor virginica
## [85] virginica  virginica  setosa   versicolor virginica virginica
## [91] virginica  versicolor virginica  versicolor virginica  virginica
## [97] setosa   setosa   virginica  versicolor virginica  setosa
## [103] virginica  setosa  versicolor versicolor versicolor virginica
## [109] setosa   setosa   versicolor virginica  versicolor versicolor
## [115] virginica  virginica  virginica  setosa  virginica  versicolor
## Levels: setosa versicolor virginica
```

Gambar 4.18 Confusion Matrix

Dari hasil confusion matrix di atas, didapat bahwa akurasi pada data train adalah 0.9583 atau 95.83% dan data test 96.67 %, artinya model_iris yang kita buat sudah cukup *appropriate-fitting* (sesuai keinginan), sehingga tidak perlu dilakukan tuning model.

Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. In machine learning, this is accomplished by selecting appropriate "hyperparameters."

Kasus DATA AGEN ASURANSI

PT XY merupakan sebuah perusahaan asuransi yang memiliki tingkat pergantian agen asuransi yang tinggi. PT XY bermaksud ingin memperbaiki sistem rekruitmen agen dengan melakukan prediksi *Total Omset* terhadap 100 calon agen asuransi berdasarkan data klasifikasi 1000 agen asuransi yang dimiliki. Hal ini dilakukan dalam upaya untuk menjaga kestabilan kinerja tiap agen dan mengurangi tingkat pergantian agen dengan hanya merekrut agen baru yang diprediksi memiliki kinerja yang baik (Class A).Class of total omset terbagi menjadi dua kelas yaitu Class A (omset > IDR 50000000), Class B.

Note:

- Data yang digunakan adalah *data_agen_asuransi.xlsx*
- Terdapat 11 variabel prediktor dan 1 kelas variabel target

Penyelesaian:

Import data agen asuransi

```
library(readxl)library(dplyr)
Agen_Asuransi <- read_excel("C:/BIBIB/MTI Data Mining/Agen Asuransi.xlsx")Asuransi <- Agen_Asuransi%>%
select(-Agent_ID)
Asuransi
## # A tibble: 1,000 x 12

##    Sex      Age Apply_Position Marital_Status Number_of_Dependa~
      Academic_Level
##    <chr>  <dbl> <chr>          <chr>           <dbl> <chr>
##                                                      5 Diploma
1 Female
50 IC
Divorced
##  2 Female   48 IC             Divorced                0 Diploma
##  3 Male     49 IC             Divorced                3 Bachelor
##  4 Male     37 EC             Married                 5 Senior High Sc~
```

```
##    5 Female    26 IC          Single              3 Diploma
##    6 Male      42 IC          Married             0 Bachelor
##    7 Male      31 EC          Single              3 Bachelor
##    8 Male      36 EC          Single              1 Bachelor
##    9 Female    52 ME          Single              3 Diploma
## 10 Female     46 ME          Divorced            0 Master
## # ... with 990 more rows, and 6 more variables: Previous_Job_Nature
<chr>,
## # Employment_Status <chr>, Past_Annual_Income <dbl>,
## # Working_Experience <dbl>, Management_Experience <dbl>, ## #
Class_of_Total_Omset <chr>
summary(Asuransi)
##        Sex          Age            Apply_Position      Marital_Status
## Length:1000 Min. :25.0   Length:1000          Length:1000
## Class :character 1st Qu.:34.0 Class :character Class :character
## Mode :character Median :43.0 Mode :character Mode :character
##          Mean :42.9
##          3rd Qu.:52.0
##          Max. :60.0
## Number_of_Dependants Academic_Level Previous_Job_Nature Employment_Status
## Min. :0.000  Length:1000          Length:1000           Length:1000
## 1st Qu.:1.000 Class :character     Class :character      Class :character
## Median :2.000 Mode :character     Mode  :character      Mode :character
## Mean  :2.465
## 3rd Qu.:4.000
##  Max.  :5.000
##  Past_Annual_Income  Working_Experience Management_Experience
##  Min.   : 50220013   Min.   :0.000      Min. :0.000
##  1st Qu.: 90316706 1st Qu.:0.000       1st Qu.:0.000
##  Median :124855072 Median :2.000       Median :0.000
##  Mean   :125011844  Mean  :1.512 Mean :0.496 3rd
## 3rd Qu.:163092544 3rd Qu.:3.000 Qu.:1.000
##  Max.  :199986029 Max.  :3.000   Max. :1.000
```

```
Max.  :199986029 Max.  :3.000
##  Class_of_Total_Omset
##  Length:1000
## Class :character
##  Mode  :character
##
##
##
```

Mengubah data kategorikal yang belum terbaca menjadi jenis factor dalam R.

```
##   1st Qu.:1.000      Diploma          :253      Management    :248
##   Median :2.000      Master           :250      Profesional   :253
##   Mean :2.4653rd     Senior High School:258 Sales/servicing   :258
##   Qu.:4.000 Max.
##   :5.000
##
##   Employment_Status Past_Annual_Income  Working_Experience
##   Employed  :487 Min.  : 50220013   Min.   :0.000
##   Unemployed:513          1st Qu.: 90316706  1st Qu.:0.000
##   Median :124855072  Median :2.000
##   Mean  :125011844   Mean   :1.512
##   3rd Qu.:163092544  3rd Qu.:3.000
##   Max.  :199986029   Max.   :3.000
##   Management_Experience Class_of_Total_Omset
##   Min.   :0.000     Class A:352
##   1st Qu.:0.000     Class B:648
##   Median :0.000
##   Mean  :0.496
## 3rd Qu.:1.000
## Max. :1.000
```

Gambar 4.19 Data Kategorikal

Semua data kategorikal sudah terbaca.

Membagi ke dalam data training dan data testing Cross validation

Kita akan melakukan cross validation (membagi data menjadi bagian-bagian tertentu). Pada kesempatan kali ini, kita membagi dtaa menjadi 2, yakni data train dan test dengan proporsi data train adalah 80% dan data test 20%.

```
set.seed(100) #Set the seed of R's random number generator, which is useful for creating simulations or random objects that can be reproduced
RNGkind(sample.kind = "Rejection") #sample.kind can be "Rounding" or "Rejection", or partial matches to these. The former was the default in versions prior to 3.6.0: it made sample noticeably non-uniform on large populations, and should only be used for reproduction of old results.
idx_Asuransi <- sample(nrow(Asuransi_fix), nrow(Asuransi_fix)*0.8) #membagi 80% dan 20%
train_Asuransi <- Asuransi_fix[idx_Asuransi,] test_Asuransi <- Asuransi_fix[-idx_Asuransi,]
```

```
Melihat data training
head(train_Asuransi)
## # A tibble: 6 x 12
##   Sex Age Apply_Position Marital_Status Number_of_Dependants Academic_Level
##  <fct> <dbl> <fct>          <fct>          <dbl> <fct>
## 1 Male    37 ME           Married            1 Bachelor
## 2 Male    30 ME           Single             2 Bachelor
## 3 Male    55 ME           Divorced           3 Master
## 4 Female  49 EC           Divorced           0 Senior High S~
## 5 Female  59 EC           Divorced           5 Diploma
## 6 Male    44 ME           Divorced           2 Bachelor
## # ... with 6 more variables: Previous_Job_Nature <fct>,
## #   Employment_Status <fct>, Past_Annual_Income <dbl>,
## # Working_Experience <dbl>, Management_Experience <dbl>,## #
Class_of_Total_Omset <fct>
```

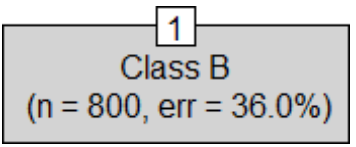
```
Melihat data testing

head(test_Asuransi)
## # A tibble: 6 x 12
##   Sex Age Apply_Position Marital_Status Number_of_Dependants
Academic_Level
##  <fct>  <dbl> <fct>          <fct>          <dbl> <fct>
## 1 Male      49 IC           Divorced       3 Bachelor
## 2 Female    26 IC           Single         3 Diploma
## 3 Male      36 EC           Single         1 Bachelor
## 4 Female    55 IC           Single         0 Master
## 5 Female    51 ME           Married        3 Master
## 6 Female    56 IC           Single         3 Bachelor
## # ... with 6 more variables: Previous_Job_Nature <fct>,
## # Employment_Status <fct>, Past_Annual_Income <dbl>,
## # Working_Experience <dbl>, Management_Experience <dbl>,
## # Class_of_Total_Omset <fct>
library(caret) # mengevaluasi model dengan confusion matrix
library(partykit) # pemodelan decision tree #PEMBUATAN MODEL
model_Asuransi <- ctree(formula = Class_of_Total_Omset ~ ., data =
train_Asuransi)
```

Gambar 4.20 Data Testing

Visualisasi model

Untuk membuat visualisasi dari model yang sudah dibuat, kita bisa menggunakan fungsi plot().

```
plot(model_Asuransi, type = "simple")
```



Mencoba dengan bentuk tree yang lain

```
library(rpart)
model.Asuransi <- rpart(Class_of_Total_Omset~., data = train_Asuransi, method = 'class') summary(model.Asuransi)


## Call:
## rpart(formula = Class_of_Total_Omset ~ ., data = train_Asuransi,
##     method = "class")
## n= 800 ##
##     CP nsplit rel error xerror     xstd
## 1 0.01085069      0 1.0000000 1.000000 0.04714045
## 2 0.01041667     14 0.8229167 1.041667 0.04754536
## 3 0.01000000     16 0.8020833 1.069444 0.04778817 ##
## Variable importance
##     Past_Annual_Income                    Age            Apply_Position
##                     36                     17                         11
## Management_Experience     Number_of_Dependants        Working_Experience
##                     11                      9                          5
##     Marital_Status      Academic_Level      Previous_Job_Nature
##                     4                      4                          2
##
## Node number 1: 800 observations,    complexity param=0.01085069
## predicted class=Class B expected loss=0.36 P(node) =1
##     class counts: 288     512
##     probabilities: 0.360 0.640
## left son=2 (29 obs) right son=3 (771 obs)
## Primary splits:
##     Past_Annual_Income < 194884200 to the right, improve=2.2121630, (0 missing)
##     Apply_Position splits as  LRR, improve=1.9695570, (0 missing)
##     Number_of_Dependants < 0.5 to the right, improve=1.5318400, (0 missing)
##     Management_Experience < 0.5 to the right, improve=1.3364700, (0 missing)
```

Age < 57.5 to the right, improve=0.9957071, (0 missing)

##

Node number 2: 29 observations

predicted class=Class A expected loss=0.4482759 P(node) =0.03625

class counts: 16 13

probabilities: 0.552 0.448

##

Node number 3: 771 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.3527886 P(node) =0.96375

class counts: 272 499

probabilities: 0.353 0.647

left son=6 (647 obs) right son=7 (124 obs) ## Primary splits:

Number_of_Dependants < 0.5 to the right, improve=2.651688, (0 missing)

Past_Annual_Income <134522800 to the left, improve=2.481015, (0 missing)

Apply_Position splits as LRR, improve=1.940529, (0 missing)

Previous_Job_Nature splits as LRLL, improve=1.227839, (0 missing)

Management_Experience < 0.5 to the right, improve=1.074544, (0 missing)

Surrogate splits:

Past_Annual_Income < 193992000 to the left, agree=0.842, adj=0.016, (0 split)

##

Node number 6: 647 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.3709428 P(node) =0.80875

class counts: 240 407

probabilities: 0.371 0.629

left son=12 (385 obs) right son=13 (262 obs) ## Primary splits:

Past_Annual_Income < 133926300 to the left, improve=2.2308240, (0 missing)

Previous_Job_Nature splits as LRLL, improve=1.9176560, (0 missing)

```
##       Apply_Position  splits as LRR, improve=1.6904240, (0 missing)
##       Age      < 50.5   to the left, improve=1.0768930, (0 missing)
##       Marital_Status splits as RLR, improve=0.8542169, (0 missing)
##
## Node number 7: 124 observations,     complexity param=0.01085069
## predicted class=Class B expected loss=0.2580645 P(node) =0.155
##       class counts:     32        92
##       probabilities: 0.258 0.742
## left son=14 (10 obs) right son=15 (114 obs)
## Primary splits:
##       Age      < 58.5   to the right, improve=4.2487830, (0 missing)
##       Working_Experience < 0.5 to the left, improve=1.3763440, (0
missing)
##       Management_Experience < 0.5  to the right, improve=1.1700540,
(0 missing)
##       Academic_Levelsplits as  RRRL, improve=0.8706598, (0 missing)
##       Employment_Status     splits as LR, improve=0.7030383, (0
missing)
##
## Node number 12: 385 observations,    complexity param=0.01085069
## predicted class=Class B expected loss=0.4051948 P(node) =0.48125
##       class counts:  156 229
##       probabilities: 0.405 0.595
## left son=24 (13 obs) right son=25 (372 obs)
## Primary splits:
##       Past_Annual_Income < 131080900 to the right,
improve=3.5659870, (0 missing)
##       Age      < 58.5   to the left,  improve=2.4446220, (0 missing)
##       Previous_Job_Nature splits as LRLL, improve=1.4578270, (0
missing)
##       Academic_Level          splits as LLRR, improve=0.7050443, (0
missing)
##       Apply_Position splits as LLR, improve=0.6884744, (0 missing)
##
```

Node number 13: 262 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.3206107 P(node) =0.3275

class counts: 84 178

probabilities: 0.321 0.679

left son=26 (93 obs) right son=27 (169 obs)

Primary splits:

Apply_Position splits as LRR, improve=2.811579, (0 missing)

Sex splits as RL, improve=2.075761, (0 missing)

Age < 58.5 to the right, improve=1.994112, (0 missing)

Number_of_Dependants < 1.5 to the left, improve=1.865468, (0 missing)

Past_Annual_Income < 137608300 to the right, improve=1.084773, (0 missing)

Surrogate splits:

Past_Annual_Income < 187352800 to the right, agree=0.656, adj=0.032, (0 split)

Age < 25.5 to the left, agree=0.649, adj=0.011, (0 split)

##

Node number 14: 10 observations

predicted class=Class A expected loss=0.3 P(node) =0.0125

class counts: 7 3

probabilities: 0.700 0.300

##

Node number 15: 114 observations

predicted class=Class B expected loss=0.2192982 P(node) =0.1425

class counts: 25 89

probabilities: 0.219 0.781

##

Node number 24: 13 observations

predicted class=Class A expected loss=0.2307692 P(node) =0.01625

class counts: 10 3

probabilities: 0.769 0.231

##

Node number 25: 372 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.3924731 P(node) =0.465

class counts: 146 226

probabilities: 0.392 0.608

left son=50 (354 obs) right son=51 (18 obs) ## Primary splits:

Age < 58.5 to the left, improve=2.9948360, (0 missing)

Past_Annual_Income < 77029610 to the right, improve=2.3750900, (0 missing)

Previous_Job_Nature splits as LRLL, improve=0.9128943, (0 missing)

Academic_Level splits as LLRR, improve=0.6674851, (0 missing)

Marital_Status splits as RLR, improve=0.5882624, (0 missing)

##

Node number 26: 93 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.4193548 P(node) =0.11625

class counts: 39 54

probabilities: 0.419 0.581

left son=52 (21 obs) right son=53 (72 obs)

Primary splits:

Number_of_Dependants < 1.5 to the left, improve=2.163338, (0 missing)

Age < 29.5 to the right, improve=1.721092, (0 missing)

Past_Annual_Income < 174767600 to the left, improve=1.674938, (0 missing) ## Working_Experience < 0.5 to the left, improve=1.352676, (0 missing)

Marital_Status splits as LLR, improve=1.152275, (0 missing)

Surrogate splits:

Past_Annual_Income < 191499800 to the right, agree=0.785, adj=0.048, (0 split)

##

Node number 27: 169 observations

predicted class=Class B expected loss=0.2662722 P(node) =0.21125

class counts: 45 124

probabilities: 0.266 0.734

##

Node number 50: 354 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.4067797 P(node) =0.4425

class counts: 144 210

probabilities: 0.407 0.593

left son=100 (228 obs) right son=101 (126 obs)

Primary splits:

Past_Annual_Income < 76556660 to the right, improve=2.5914010, (0 missing)

Age < 54.5 to the right, improve=1.7921290, (0 missing)

Previous_Job_Nature splits as LRLL, improve=1.0996270, (0 missing)

Academic_Levelsplits as LLRR, improve=0.9930202, (0 missing)

Number_of_Dependants < 2.5 to the right, improve=0.6314787, (0 missing)

##

Node number 51: 18 observations

predicted class=Class B expected loss=0.1111111 P(node) =0.0225

class counts: 2 16

probabilities: 0.111 0.889

Node number 52: 21 observations, complexity param=0.01085069

predicted class=Class A expected loss=0.3809524 P(node) =0.02625

class counts: 13 8

probabilities: 0.619 0.381

left son=104 (13 obs) right son=105 (8 obs)

Primary splits:

Management_Experience < 0.5 to the right, improve=6.3086080, (0 missing)

Marital_Status splits as RLR, improve=1.6932230, (0 missing)

Age < 48 to the left, improve=0.9603175, (0 missing)

Past_Annual_Income < 180223800 to the right, improve=0.7936508, (0 missing)

Academic_Levelsplits as RRLR, improve=0.4432234, (0 missing)

Surrogate splits:

Marital_Status splits as LLR, agree=0.762, adj=0.375, (0 split)

Age < 25.5 to the right, agree=0.714, adj=0.250, (0 split)

Working_Experience < 1.5 to the right, agree=0.714, adj=0.250, (0 split)

Previous_Job_Nature splits as LRLL, agree=0.667, adj=0.125, (0 split)

##

Node number 53: 72 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.3611111 P(node) =0.09

class counts: 26 46

probabilities: 0.361 0.639

left son=106 (45 obs) right son=107 (27 obs) ## Primary splits:

Past_Annual_Income < 175263600 to the left, improve=2.6740740, (0 missing)

Working_Experience < 0.5 to the left, improve=1.9760680, (0 missing)

Age < 37.5 to the right, improve=1.3827570, (0 missing)

Sex splits as RL, improve=1.1313130, (0 missing)

Number_of_Dependants < 4.5 to the right, improve=0.7936508, (0 missing)

Surrogate splits:

Age < 26.5 to the right, agree=0.681, adj=0.148, (0 split)

##

Node number 100: 228 observations, complexity param=0.01085069

predicted class=Class B expected loss=0.4517544 P(node) =0.285

class counts: 103 125

probabilities: 0.452 0.548

left son=200 (92 obs) right son=201 (136 obs)

Primary splits:

Past_Annual_Income < 99460280 to the left, improve=3.2467810, (0 missing)

Age < 57.5 to the right, improve=1.1700780, (0 missing)

Previous_Job_Nature splits as LRLL, improve=1.0050300, (0

```
missing)
##       Academic_Level  splits as LLRR, improve=0.9894439, (0 missing)
##       Apply_Position splits as LLR, improve=0.6377569, (0 missing)
## Surrogate splits:
##       Number_of_Dependants < 1.5    to the left, agree=0.618, adj=0.054,
(0 split)
 ##       Age      < 57.5    to the right, agree=0.605, adj=0.022, (0 split)
##
## Node number 101: 126 observations
## predicted class=Class B expected loss=0.3253968 P(node) =0.1575
##       class counts:     41        85
##       probabilities: 0.325 0.675 ##
## Node number 104: 13 observations
## predicted class=Class A expected loss=0.07692308 P(node) =0.01625
##       class counts:     12        1
##       probabilities: 0.923 0.077
##
## Node number 105: 8 observations
## predicted class=Class B expected loss=0.125 P(node) =0.01
##       class counts:     1         7
##       probabilities: 0.125 0.875
##
## Node number 106: 45 observations,    complexity param=0.01085069
## predicted class=Class B expected loss=0.4666667 P(node) =0.05625
##       class counts:     21        24
##       probabilities: 0.467 0.533
## left son=212 (13 obs) right son=213 (32 obs)
## Primary splits:
##       Past_Annual_Income < 162379000 to the right,
improve=3.3471150, (0 missing)
##       Working_Experience < 0.5        to the left, improve=1.9775400, (0
missing)
##       Number_of_Dependants < 4.5    to the right, improve=1.5621620,
(0 missing)
```

Age < 48.5 to the left, improve=1.5363640, (0 missing)

Marital_Status splits as LLR, improve=0.3090909, (0 missing)

##

Node number 107: 27 observations

predicted class=Class B expected loss=0.1851852 P(node) =0.03375

class counts: 5 22

probabilities: 0.185 0.815

Node number 200: 92 observations, complexity param=0.01085069

predicted class=Class A expected loss=0.4456522 P(node) =0.115

class counts: 51 41

probabilities: 0.554 0.446

left son=400 (63 obs) right son=401 (29 obs)

Primary splits:

Apply_Position splits as LLR, improve=3.718153, (0 missing)

Previous_Job_Nature splits as LRLL, improve=2.186251, (0 missing) ## Academic_Levelsplits as RLRR, improve=1.883984, (0 missing)

Past_Annual_Income < 97830820 to the right, improve=1.801760, (0 missing)

Age < 57.5 to the right, improve=1.389295, (0 missing)

Surrogate splits:

Previous_Job_Nature splits as LRLL, agree=0.728, adj=0.138, (0 split)

Past_Annual_Income < 80434500 to the right, agree=0.696, adj=0.034, (0 split)

##

Node number 201: 136 observations, complexity param=0.01041667

predicted class=Class B expected loss=0.3823529 P(node) =0.17

class counts: 52 84

probabilities: 0.382 0.618

left son=402 (33 obs) right son=403 (103 obs)

Primary splits:

Academic_Levelsplits as LRRR, improve=2.3182600, (0 missing)

Past_Annual_Income < 128031200 to the left, improve=0.9625668,

(0 missing)

Age < 47.5 to the right, improve=0.7686275, (0 missing)

Working_Experience < 1.5 to the right, improve=0.7109886, (0 missing)

Previous_Job_Nature splits as RRLR, improve=0.7058824, (0 missing)

Surrogate splits:

Age < 56.5 to the right, agree=0.765, adj=0.03, (0 split)

##

Node number 212: 13 observations

predicted class=Class A expected loss=0.2307692 P(node) =0.01625

class counts: 10 3

probabilities: 0.769 0.231

##

Node number 213: 32 observations

predicted class=Class B expected loss=0.34375 P(node) =0.04

class counts: 11 21

probabilities: 0.344 0.656

##

Node number 400: 63 observations

predicted class=Class A expected loss=0.3492063 P(node) =0.07875

class counts: 41 22

probabilities: 0.651 0.349

##

Node number 401: 29 observations

predicted class=Class B expected loss=0.3448276 P(node) =0.03625

class counts: 10 19

probabilities: 0.345 0.655

##

Node number 402: 33 observations, complexity param=0.01041667

predicted class=Class A expected loss=0.4545455 P(node) =0.04125

class counts: 18 15

probabilities: 0.545 0.455

left son=804 (20 obs) right son=805 (13 obs)

Primary splits:

Working_Experience < 2.5 to the left, improve=1.1097900, (0 missing)

Age < 48.5 to the right, improve=1.0909090, (0 missing)

Apply_Position splits as RRL, improve=0.6853755, (0 missing)

Marital_Status splits as RRL, improve=0.6853755, (0 missing)

Past_Annual_Income < 112923600 to the left, improve=0.6643882, (0 missing)

Surrogate splits:

Past_Annual_Income < 108592700 to the right, agree=0.727, adj=0.308, (0 split)

Age < 38.5 to the right, agree=0.697, adj=0.231, (0 split)

Number_of_Dependants < 4.5 to the left, agree=0.667, adj=0.154, (0 split)

Marital_Status splits as RLL, agree=0.636, adj=0.077, (0 split)

Previous_Job_Nature splits as RLLL, agree=0.636, adj=0.077, (0 split)

##

Node number 403: 103 observations

predicted class=Class B expected loss=0.3300971 P(node) =0.12875

class counts: 34 69

probabilities: 0.330 0.670

##

Node number 804: 20 observations

predicted class=Class A expected loss=0.35 P(node) =0.025

class counts: 13 7

probabilities: 0.650 0.350

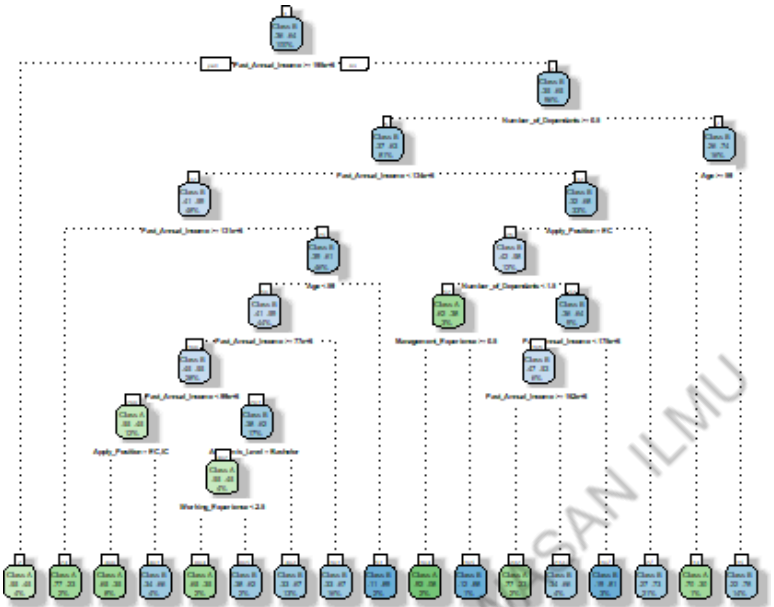
##

Node number 805: 13 observations

predicted class=Class B expected loss=0.3846154 P(node) =0.01625

class counts: 5 8

probabilities: 0.385 0.615 library(rattle)



Rattle 2022-Apr-14 22:01:08 anik nur habyba

model.Asuransi$variable.importance

Past_Annual_Income Age Apply_Position

20.574513 9.644101 6.529732

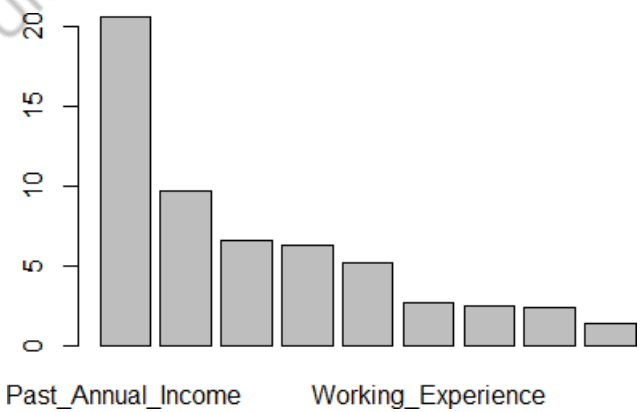
Management_Experience Number_of_Dependants Working_Experience

6.308608 5.162219 2.686942

Marital_Status Academic_Level Previous_Job_Nature

2.451096 2.318260 1.386793

barplot(model.Asuransi$variable.importance)



```
variable <- data.frame(model.Asuransi$variable.importance) variable

## model.Asuransi.variable.importance
## Past_Annual_Income        20.574513
## Age  9.644101
## Apply_Position            6.529732
## Management_Experience     6.308608
## Number_of_Dependants      5.162219
## Working_Experience        2.686942
## Marital_Status            2.451096
## Academic_Level            2.318260
## Previous_Job_Nature       1.386793
```

Prediksi Data Testing Selanjutnya memprediksi data testing untuk melihat akurasi.

```
# prediksi testing
pred_Asuransi <- predict(model.Asuransi, newdata=test_Asuransi, type="class")
# Confusion matrix
table(pred_Asuransi, test_Asuransi$Class_of_Total_Omset)
##
## pred_Asuransi Class A Class B
##    Class A       15      37
##    Class B       49      99
```

Selanjutnya memprediksi data testing untuk melihat akurasi.

```
library(caret)
# prediksi testing
pred_Asuransi <- predict(model.Asuransi, newdata=test_Asuransi, type="class")
# Confusion matrix
confusionMatrix(data=pred_Asuransi, reference=test_Asuransi$Class_of_Total_Omset)
## Confusion Matrix and Statistics ##
##          Reference
```

```
## Prediction Class A Class B
## Kelas A       15       37
## Kelas B       49       99
##                Accuracy : 0.57
##                  95% CI : (0.4983, 0.6396)
## No Information Rate : 0.68
## P-Value [Acc > NIR] : 0.9996
##
## Kappa : -0.0397
## Mcnemar's Test P-Value : 0.2356
##
##             Sensitivity : 0.2344
##             Specificity : 0.7279
##          Pos Pred Value : 0.2885
##          Neg Pred Value : 0.6689
##              Prevalence : 0.3200
##          Detection Rate : 0.0750
## Detection Prevalence : 0.2600
##    Balanced Accuracy : 0.4812
```

Gambar 4.21 Prediksi Data Testing

Hasil akurasi masih rendah yaitu 0.57, dapat diambil kesimpulan dari indikator yang lain. Hasil akurasi model sangat dipengaruhi oleh kualitas data.

MELAKUKAN PREDIKSI KINERJA 100 KARYAWAN BARU DENGAN MODEL YANG DIHASILKAN

```
library(readxl) library(dplyr)
Agen_Asuransi <- read_excel("C:/BIBIB/MTI Data Mining/Agen_testing.xlsx") Agen_baru <- Agen_Asuransi%>% select(-Agent_ID)
Agen_baru
## # A tibble: 100 x 12
##      Sex     Age Apply_Position Marital_Status Number_of_Dependa~ Academic_Level ##    <chr> <dbl> <chr>          <chr>    <dbl>          <chr>
```

```
## 1 Male      32 ME  Divorced      5 Master
## 2 Male      31 IC  Divorced      4 Diploma
## 3 Male      35 EC  Divorced      3 Diploma
## 4 Female    29 IC  Single        3 Diploma
## 5 Male      29 ME  Divorced      2 Diploma
## 6 Female    29 EC  Single        3 Master
## 7 Female    26 IC  Divorced      0 Senior High Sc~
## 8 Female    33 EC  Married       2 Master
## 9 Male      38 IC  Divorced      4 Bachelor
## 10 Female   28 EC  Divorced      5 Diploma
## # ... with 90 more rows, and 6 more variables: Previous_Job_Nature
<chr>,
## # Employment_Status <chr>, Past_Annual_Income <dbl>,
## # Working_Experience <dbl>, Management_Experience <dbl>,
## #  Class_of_Total_Omset <chr>
summary(Agen_baru)
##    Sex                Age           Apply_Position      Marital_Status
## Length:100      Min. :25.00    Length:100          Length:100
## Class :character 1st Qu.:28.75 Class :character Class :character
## Mode :character Median :32.00 Mode :character Mode :character
##    Mean :32.49
##    3rd Qu.:36.25
##    Max. :40.00
## Number_of_Dependants Academic_Level Previous_Job_Nature
Employment_Status
## Min. :0.0      Length:100        Length:100        Length:100
## 1st Qu.:1.0    Class :character Class :character Class :character
## Median :3.0    Mode :character Mode  :character Mode :character
## Mean  :2.7
## 3rd Qu.:4.0 ## Max. :5.0
## Past_Annual_Income Working_Experience Management_Experience
## Min. : 54375961 Min. :0.0     Min. :0.00
## 1st Qu.: 92492899 1st Qu.:0.0         1st Qu.:0.00
```

```
## Median :125939391 Median :1.0       Median :1.00
## Mean   :124174729   Mean   :1.4     Mean   :0.58
## 3rd Qu.:159802340   3rd Qu.:2.0     3rd Qu.:1.00
## Max.  :199814907 Max. :3.0 Max. :1.00
## Class_of_Total_Omset
## Length:100
## Class :character
## Mode :character
##
## ##
Agen_baru <- Agen_baru%>% mutate_if(is.character, as.factor)
Agen_baru
## # A tibble: 100 x 12
##  SexAge Apply_Position Marital_Status Number_of_Dependa~
Academic_Level
##  <fct> <dbl> <fct>    <fct>    <dbl> <fct>
## 1 Male        32 ME  Divorced        5 Master
## 2 Male        31 IC  Divorced        4 Diploma
## 3 Male        35 EC  Divorced        3 Diploma
## 4 Female      29 IC  Single          3 Diploma
## 5 Male        29 ME  Divorced        2 Diploma
## 6 Female      29 EC  Single          3 Master
## 7 Female      26 IC  Divorced        0 Senior High Sc~
## 8 Female      33 EC  Married         2 Master
## 9 Male        38 IC  Divorced        4 Bachelor
## 10 Female     28 EC  Divorced        5 Diploma
## # ... with 90 more rows, and 6 more variables: Previous_Job_Nature
<fct>,
## # Employment_Status <fct>, Past_Annual_Income <dbl>,
## # Working_Experience <dbl>, Management_Experience <dbl>,
## #  Class_of_Total_Omset <fct>
```

Selanjutnya model yang telah dibuat digunakan untuk memprediksi data baru

```
library(caret)
# prediksi testing
pred_agen_baru <- predict(model.Asuransi, newdata=Agen_baru, type="-class") summary(pred_agen_baru)
## Class A Class B
##     22      78
```

Dari total 100 karyawan baru hanya 22 orang yang masuk kinerja A. Hal ini tentunya dihasilkan dengan akurasi model sebesar 57%. Perlu dipertimbangkan untuk memperbaiki model terlebih dahulu sebelum digunakan untuk pengambilan keputusan.

```
library(dplyr)
pred_result <- cbind(Agen_baru, pred_agen_baru)head(pred_result)
##   Sex Age Apply_Position Marital_Status Number_of_Dependants Academic_Level
## 1 Male  32 ME Divorced 5 Master
## 2 Male  31 IC Divorced 4 Diploma
## 3 Male  35 EC Divorced 3 Diploma
## 4 Female 29 IC Single 3 Diploma
## 5 Male  29 ME Divorced 2 Diploma
##6 Female 29 EC Single 3 Master
## Previous_Job_Nature Employment_Status Past_Annual_Income Working_Experience
## 1 Clerical/technician Employed 141607668 3
## 2 Clerical/technician Employed 124271414 0
## 3 Management Employed 149727062 1
## 4 Profesional Unemployed 66263432 1
## 5 Profesional Unemployed 99721535 2
## 6 Clerical/ technician Employed 137900554 2
```

Untuk melihat keseluruhan hasil prediksi 100 karyawan dapat dilakukan dengan berikut:

```
##  Management_Experience Class_of_Total_Omset pred_agen_baru
## 11 - Class B
## 21 - Class B
```

```
## 30          -       Class B
## 41          -       Class B
## 50          -       Class B
## 60          -       Class B
view(pred_result)
```

4.8. Metode Klasifikasi *Data Mining Naïve bayes*

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Teorema tersebut dikombinasikan dengan *Naive* di mana diasumsikan kondisi antar atribut saling bebas. Algoritma *Naive Bayes* merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.

Algoritma *Naive Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema *Bayes*. Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan kondisi antar atribut saling bebas. Untuk menjelaskan teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Ciri utama dari Naïve Bayes *Classifier* ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian, *Naive Bayes Classifier* bekerja sangat baik dibanding dengan model classifier lainnya. (Zhang,2004)

Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai variabel independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

Tahapan Naïve Bayes

Tahapan dari proses algoritma Naive Bayes adalah:

1. Menghitung jumlah kelas / label.
2. Menghitung Jumlah Kasus Per Kelas
3. Kalikan Semua Variable Kelas
4. Bandingkan Hasil Per Kelas

Dasar dari Naïve Bayes yang dipakai dalam pemrograman adalah rumus Bayes:

$P(A|B) = (P(B|A) * P(A))/P(B)$ .....(1)

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$P(Ci|D) = (P(D|Ci)*P(Ci)) / P(D)$ ..... (2)

Naïve Bayes Classifier atau bisa disebut sebagai Multinomial Naïve Bayes merupakan model penyederhanaan dari Metoda Bayes yang cocok dalam pengklasifikasian teks atau dokumen.

Persamaannya adalah: $V_{MAP} = \arg\max P(Vj \mid a_1, a_2, ...a_n)$ .....(3)

Menurut persamaan (3), maka persamaan (1) dapat ditulis:

$$V_{MAP} = \arg\max_{vj \in V} \frac{P(a_1, a_2, ... a_n|v_j)P(v_j)}{P(a_1, a_2, ..., a_n)} \text{.....(4)}$$

$P(a_1, a_2, ...a_n)$ konstan, sehingga dapat dihilangkan menjadi

$\arg\max_{vj \in V} = P(a_1, a_2, ... a_n|v_j)P(v_j)$ ..... (5)

4.9. Studi Kasus Naïve Bayes

Metode *data mining* untuk mengklasifikasi seluruh peristiwa *non productive time* yang terdapat pada daily drilling report *system perbaikan kualitas* pada *Integrated project Improvement drilling* PT.XYZ.

Sumber Data

Data yang di gunakan dalam metode ini merupakan data *non productive time* yang terjadi pada projek pengeboran panas bumi X yang berlangsung dari bulan mei 2014 sampai bulan Mei 2018.

Pre-Processing Data

Klasifikasi data untuk pembuatan prediksi data menggunakan metode *Naïve Bayes* ditentukan bedasarkan persetujuan yang sudah disetujui oleh pihak PT.XYZ dan pihak *client* PT.XYZ adalah sebagai berikut.

a) *Class Operation*

Class operation pada proyek di klasifikasikan menjadi tiga yaitu P jika operasi proyek berjalan sesuai rencana, U jika operasi proyek tidak di jalankan sesuai dengan rencana di karenakan beberapa faktor dari kondisi lingkungan pada proyek pengeboran X dan NPT jika operasi tersebut menglami *non productive time.*

Tabel 4.19 klasifikasi *Class Operation*

Class Operation	Pengertian
P	Operasi dijalankan sesuai dengan rencana
U	Operasi dijalankan tidak sesuai dengan rencana
NPT	Operasi mengalami *non productive time*

b) Tingkat *non productive time*

Tingkat *non productive time* pada proyek di klasifikasi kan menjadi tiga yaitu "rendah" jika waktu *non productive time* lebih kecil atau sama dengan 2,5 jam, "Sedang" jika *waktu non productive time* berlangsung lebih kecil atau sama dengan 3 jam, dan "Tinggi" jika waktu *non productive time* berlangsung lebih dari 5 jam.

Tabel 4.20 Klasifikasi tingkat *Non Productive Time*

Tingkat Non productive time	Pengertian
Rendah	Operasi NPT< 2.5 jam
Sedang	Operasi NPT<= 3 jam
Tinggi	Operasi NPT > 5 jam

c) Batas toleransi NPT

Batas toleransi NPT diklasifikasikan menjadi dua yaitu "*di dalam batas toleransi*" dan "*di luar batas toleransi*"

Tabel 4.21 Klasifikasi tingkat toleransi yang diberikan *client*

Tingkat Toleransi NPT	Pengertian
di dalam batas toleransi	NPT < 5 Jam
di luar batas toleransi	NPT > 5 Jam

Tingkat toleransi yang ditetapkan client tidak seluruhnya diukur bedasarkan durasi NPT yang berlangsung, terdapat pula beberapa variabel untuk mengukur tingkat toleransi yang ditetapkan *client*, seperti *HSE performance, total well cost, subsurface objective*, dan lain nya, untuk penelitian awal ini, variabel yang digunakan hanya durasi tiap operasi NPT yang terjadi dalam proses operasi pengeboran.

Pembuatan prediksi terhadap data *non productive time* menggunakkan metode *naïve bayes*

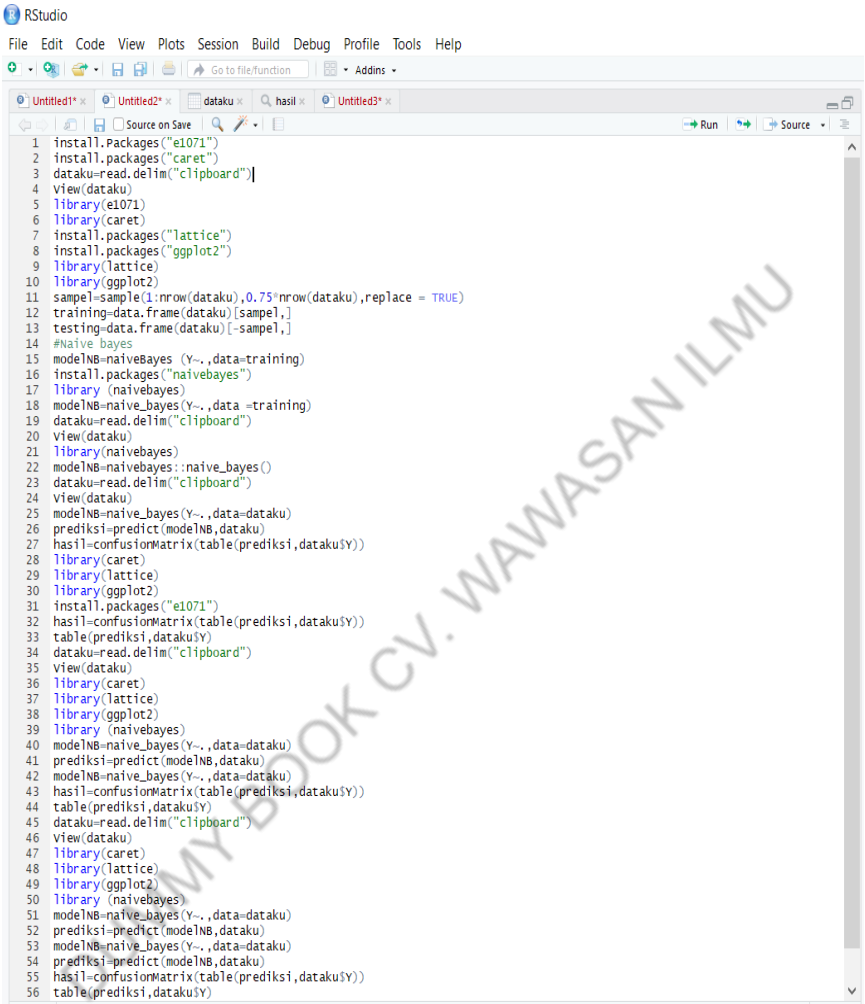
Dalam membuat prediksi untuk menentukan permasalahan apa saja yang berada diluar batas toleransi yang telah ditentukan oleh *client* terhadap kinerja proyek PT.XYZ, metode naïve bayes dapat digunakan untuk mendapatkan prediksi sesuai dengan output yang diinginkan PT.XYZ.

Menerapkan metode naïve bayes dapat dilakukan dengan menggunakan software R-Studio, R Studio adalah bahasa pemrograman dan sistem perangkat lunak yang dirancang khusus untuk mengerjakan segala hal terkait komputasi statistik.

Data yang akan diolah adalah data *tracker* operasi "*non productive time"* yang berlangsung selama pelaksanaan proyek pengeboran, proyek pengeboran berlangsung selama 3 tahun yaitu dari bulan Mei tahun 2014 Agustus 2018.

Menerapkan metode naïve bayes pada software R Studio harus menginstall beberapa aplikasi *packages* pada R studio seperti packages e170, dan caret setelah menginstall aplikasi packages tersebut packages yang harus di install berikutnya adalah package lattice, ggplot2 dan naïve bayes, setelah menginstall package tersebut, langkah yang dilakukan berikutnya adalah mengidentifikasi data menggunakan command read delim, sebelum data diidentifikasi data harus di pastikan memiliki variable x dan y sehingga data dapat diidentifikasi dan metode klasifikasi prediksi naïve bayes dapat diterapkan, setelah data diperoleh aplikasi yang sudah diinstall diaktifkan fungsinya menggunakan command library lalu menggunakan command "*ModelNB=dataku~,Y"* berikut

adalah proses pemrograman metode naïve bayes yang di lakukan menggunakan software R-Studio.



Gambar 4.22 Proses Pengolahan *Data Mining*

Proses programming yang telah dilakukan akan menunjukan output yaitu berupa prediksi dari proses operasi *non productive time* yang berada di luar batas toleransi yang ditetapkan *client*.

Berikut adalah hasil output yang didapat dari proses pengolahan data menggunakan metode naïve bayes pada *software* R Studio.

Klasifikasi	Frequensi
di dalam batas toleransi	1376
di luar batas toleransi	161

Gambar 4.23 hasil output klasifikasi *Data Mining*

Bedasarkan data yang telah di peroleh, dapat di simpulkan bahwa terdapat 161 permasalahan *non productive time* yang melebihi tingkat batas toleransi yang diberikan *client*, sehingga perbaikan terhadap 161 permasalahan "*non productive time*" harus dilakukan agar kualitas proyek dapat menjadi lebih baik pada projek berikutnya.

Perbandingan dengan perhitungan manual metode naïve bayes

Setelah mendapat hasil perhitungan dari hasil aplikasi software R-studio, hasil tersebut dibandingkan dengan perhitungan manual menggunakan metode naïve bayes, perhitungan klasifikasi naïve bayes diawali dengan *pre processing* data dan mengambil sampel dari data *daily drilling program* proyek, salah satu sampel yang diambil dijadikan case study yang kemudian dilakukan perhitungan terhadap probabilitas *case study* tersebut, setelah itu probabilitas *non productive time* yang didapat diklasifikasi menjadi 2 yaitu *non productive time* yang berada di dalam batas toleransi dan di luar batas toleransi, 2 klasifikasi tersebut memiliki total probabilitas yang kemudian dijadikan nilai P, nilai P *non productive time* yang di dalam batas toleransi dan di luar batas toleransi dibandingkan kedua nilai nya, nilai P yang lebih besar adalah hasil klasifikasi dari 1 sampel *case study* yang diambil, berikut adalah perhitungan manual dan rumus yang digunakan untuk klasifikasi naïve bayes:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \qquad 4\text{-}1$$

Tabel 4.22 Perhitungan manual naïve bayes

Klasifikasi operasi	Pengertian
P	Operasi dijalankan sesuai rencana
U	Operasi dijalankan tidak sesuai rencana
NPT	Operasi mengalami *non productive time*

Tingkat *non productive time*	Pengertian
Rendah	Operasi *NPT <= 2,5 jam*
Sedang	Operasi NPT <= 3 jam
Tinggi	Operasi NPT > 5 jam

Tingkat toleransi NPT	Pengertian
didalam batas toleransi	*Client* puas jika *non productive time* tidak lebih dari 5 jam
diluar batas toleransi	*Client* puas jika *non productive time* lebih dari 5 jam

Deskripsi operasi	klasifikasi operasi	Durasi operasi	Tingkat NPT
Repair Main Top Drive System	NPT	11.00	Sedang
Make Up BHA	P	4.75	
Reaming/Back Reaming	P	2.50	
Circulate Hole Clean	P	0.50	
Tripping out of Hole DP in CH	P	1.75	
Repair Main Top Drive System	NPT	3.00	Rendah
Nipple down BOP	P	4.00	
Set slips and cut casing	P	4.00	
Nipple Up Wellhead / X-Mas Tree / BPV	P	2.50	
Nipple Up Wellhead / X-Mas Tree / BPV	P	5.00	
Nipple Up Wellhead / X-Mas Tree / BPV	P	1.50	
Emergency Response Drills	P	0.50	
Nipple Up Wellhead / X-Mas Tree / BPV	P	6.00	
Pressure Test Casing /Tubing	P	0.50	
Pressure Test Casing /Tubing	P	0.50	
Safety Meeting	P	0.50	
Nipple Up BOP	P	15.50	
Function Test BOP	P	0.50	
Pressure Test BOP/Wellhead/X-Mas Tree	P	2.00	
Pick Up / Lay Down BHA	P	5.00	
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	16.00	Sedang

Case Study			
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI

Tahap 1		
Operasi NPT	7/25	0.28
Operasi non NPT	18/25	0.72

Tahap 2		
NPT Rendah	1/7	0.14
NPT Sedang	2/7	0.29
NPT Tinggi	4/7	0.57

P (didalam batas toleransi)			
NPT Rendah	1/7	0.14	
NPT Sedang	2/7	0.29	0.43
NPT Tinggi	0/7	0.00	

P (diluar batas toleransi)			
NPT Rendah	0/7	0.00	
NPT Sedang	0/7	0.00	0.57
NPT Tinggi	4/7	0.57	

Dikarenakan P(diluar batas toleransi) > p (didalam batas toleransi) maka hasil yang di dapat adalah

Output				
Waiting on Rig (Personnel, Equipment, or Materials)	NPT	24.00	TINGGI	diluar batas toleransi

1. Pengolahan data *Naïve Bayes* pada data hasil komentar pelanggan merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema atau aturan bayes dengan asumsi independensi yang kuat pada fitur, artinya bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Atribut yang digunakan adalah *case origin, comment type, service type, dan unit to charge.* Hasil dari penelitian ini digunakan sebagai salah satu dasar pengambilan keputusan untuk menentukan kebijakan oleh pihak perusahaan PT. X. Pengolahan data *Naïve Bayes* pada data hasil komentar pelanggan terdiri dari empat tahap.

Tahap Pertama

Pada tahap pertama, dilakukan perhitungan probabilitas untuk masing-masing sub kelas pada atribut klasifikasi unit to charge yaitu

divisi yang menangani komentar pelanggan. Rumus perhitungan manual yang digunakan pada pengolahan data *Naïve Bayes* tahap pertama adalah sebagai berikut :

P (Wi|C) = ……(1)

Dimana menyatakan probabilitas muncul jika diketahui Hasil dari perhitungan manual yang digunakan pada pengolahan data *Naïve Bayes* tahap pertama dapat dilihat pada tabel 2 yang berisi rangkuman perhitungan *Naïve Bayes* rute Jakarta menuju Singapore.

Tabel 4.23 Probabilitas *unit to charge* rute Jakarta-Singapore

Unit to Charge	Jumlah	Prosentase
Customer Care	256	0.88
Branch Office	29	0.10
Customer Relation	4	0.014
Digital Business	2	0.007
Total	291	1.00

Pada tabel 2 merupakan perhitungan probabilitas masing-masing kelas unit to charger dari unit to charge rute penerbangan Jakarta menuju Singapore berdasarkan klasifikasi yang terbentuk (*prior probability*) adalah sebagai berikut :

1. C1 (Class Unit to Charge = "Customer Care") = jumlah "Customer Care" pada kolom Unit to Charge = 256/291 = 0.88.
2. C2 (Class Unit to Charge = "Branch Office") = jumlah "Branch Office" pada kolom Unit to Charge = 29/291 = 0.014.
3. C3 (Class Unit to Charge = "Customer Relation") = jumlah "Customer Relation" pada kolom Unit to Charge = 4/291 = 0.01.
4. C4 (Class Unit to Charge = "Digital Business") = jumlah "Digital Business" pada kolom Unit to Charge = 2/291 = 0.007.

Dapat disimpulkan berdasarkan hasil perhitungan manual probabilitas masing-masing kelas unit to charger dari unit to charge rute penerbangan Jakarta menuju Singapore, kelas *unit to charge* dengan probabilitas tertinggi adalah kelas *customer care* dengan nominal probabilitas 0.88 dan kelas *unit to charge* dengan probabilitas terendah adalah kelas *digital business* dengan nominal probabilitas 0.007. Berdasarkan hasil pengolahan data menggunakan software Weka, hasil jumlah masing-masing kelas pada klasifikasi sama

dengan jumlah perhitungan manual, dapat disimpulkan bahwa perhitungan manual yang ada adalah valid. Berikut tampilan gambar 2 merupakan hasil pengolahan data untuk kelas klasifikasi unit to charge menggunakan software Weka rute Jakarta menuju Singapore.

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 256   0   0   0 |   a = Customer Care
   0  29   0   0 |   b = Branch Office
   0   1   3   0 |   c = Customer Relation
   0   1   0   1 |   d = Digital Business
```

Gambar 4.24 Hasil perhitungan software Weka klasifikasi Unit to Charge rute Jakarta Singapore

Tahap Kedua

Pada tahap kedua, dilakukan perhitungan prior probabilitas untuk masing-masing hubungan atribut klasifikasi yang ada pada komentar pelanggan. Rumus perhitungan manual yang digunakan pada pengolahan data *Naïve Bayes* tahap kedua adalah sebagai berikut :

P(D|C) = ……(2)

Dimana menyatakan nilai perkalian seluruh prior probabilitas muncul jika diketahui Hasil dari perhitungan manual yang digunakan pada pengolahan data *Naïve Bayes* tahap pertama dapat dilihat pada tabel yang berisi rangkuman perhitungan *Naïve Bayes* rute Jakarta menuju Singapore.

Tabel 4.24 Perhitungan Manual Prior Probabilitas Atribut Klasifikasi Case Origin dan Unit to Charge rute Jakarta menuju Singapore

Unit to Charge / Case Origin	Branch Office	Customer Relation	Customer care	Digital Business
Call Center/Phone	0.03	0.25	0.01	0
Customer Care Email	0	0	0.02	0
Garuda Miles Email	0	0	0	0.5
Mail	0.03	0	0	0
Suggestion Form	0.86	0.25	0.92	0.5
Walk In	0.07	0	0.004	0
Web	0	0.5	0.05	0
Total	1	1	1	1

Pada tabel 4.24 merupakan perhitungan prior probabilitas prior antara masing-masing sub kelas pada klasifikasi case origin dan unit to charge rute penerbangan Jakarta menuju Singapore yaitu dengan perhitungan manual sebagai berikut:

1. P(Case Origin "*Call Center / Phone*" | Class Unit to Charge = "*Branch Office*") = 1/29 = 0.03
2. P(Case Origin "*Customer Care Email*" | Class Unit to Charge = "*Branch Office*") = 0/29 = 0
3. P(Case Origin "*PT.X Miles Email*" | Class Unit to Charge = "*Branch Office*") = 0/29 = 0
4. P(Case Origin "*Mail*" | Class Unit to Charge = "*Branch Office*") = 1/29 = 0.03
5. P(Case Origin "*Suggestion Form*" | Class Unit to Charge = "*Branch Office*") = 25/29 = 0.86
6. P(Case Origin "*Wak in*" | Class Unit to Charge = "*Branch Office*") = 2/29 = 0.07
7. P(Case Origin "*Web*" | Class Unit to Charge = "*Branch Office*") = 0/29 = 0
8. P(Case Origin "*Call Center / Phone*" | Class Unit to Charge = "*Customer Relation*") = 1/4 = 0.25
9. P(Case Origin "*Customer Care Email*" | Class Unit to Charge = "*Customer Relation*") = 0/4 = 0.25
10. P(Case Origin "*PT.X Miles Email*" | Class Unit to Charge = "*Customer Relation*") = 0/4 = 0.25
11. P(Case Origin "*Mail*" | Class Unit to Charge = "*Customer Relation*") = 0/4 = 0.25
12. P(Case Origin "*Suggestion Form*" | Class Unit to Charge = "*Customer Relation*") = 1/4 = 0.25
13. P(Case Origin "*Wak in*" | Class Unit to Charge = "*Customer Relation*") = 0/4 = 0.25
14. P(Case Origin "*Web*" | Class Unit to Charge = "*Customer Relation*") = 2/4 = 0.5
15. P(Case Origin "*Call Center / Phone*" | Class Unit to Charge = "*Customer Care*") = 2/256 = 0.01
16. P(Case Origin "*Customer Care Email*" | Class Unit to Charge = "*Customer Care*") = 5/256 = 0.02

17. P(Case Origin "*PT.X Miles Email*" | Class Unit to Charge = "*Customer Care*") = 0/256 = 0
18. P(Case Origin "*Mail*" | Class Unit to Charge = "*Customer Care*") = 0/256 = 0
19. P(Case Origin "*Suggestion Form*" | Class Unit to Charge = "*Customer Care*") = 236/256 = 0.92
20. P(Case Origin "*Wak in*" | Class Unit to Charge = "*Customer Care*") = 1/256 = 0.004
21. P(Case Origin "*Web*" | Class Unit to Charge = "*Customer Care*") = 12/256 = 0.05
22. P(Case Origin "*Call Center / Phone*" | Class Unit to Charge = "*Digital Business*") = 0/2 = 0
23. P(Case Origin "*Customer Care Email*" | Class Unit to Charge = "*Digital Business*") = 0/2 = 0
24. P(Case Origin "*PT.X Miles Email*" | Class Unit to Charge = "*Digital Business*") = 1/2 = 0.5
25. P(Case Origin "*Mail*" | Class Unit to Charge = "*Digital Business*") = 0/2 = 0
26. P(Case Origin "*Suggestion Form*" | Class Unit to Charge = "*Digital Business*") = 1/2 = 0.5
27. P(Case Origin "*Walk in*" | Class Unit to Charge = "*Digital Business*") = 0/2 = 0
28. P(Case Origin "*Web*" | Class Unit to Charge = "*Digital Business*") = 0/2 = 0

Berdasarkan hasil perhitungan tabel 3 di atas, dapat disimpulkan bahwa prior probabilitas terbesar antara kelas pada klasifikasi *case origin* dan kelas pada klasifikasi *unit to charge* pada rute Jakarta menuju Singapore adalah prior probabilitas antara *customer care* dan *suggestion form* dengan jumlah kelas *customer care* yaitu 236 dibagi dengan jumlah seluruh kelas *suggestion form* yaitu 256, sehingga hasil prior probabilitas antara keduanya adalah 0.92. Berdasarkan hasil pengolahan data menggunakan software Weka, hasil jumlah prior probabilitas terbesar adalah antara atribut klasifikasi *case origin* dan *unit to charge* adalah sub kelas *customer care* dan *suggestion form* yaitu sebesar 237, maka dapat disimpulkan bahwa perhitungan manual yang telah dilakukan adalah valid. Berikut tampilan gambar 4 merupakan hasil pengolahan data untuk nilai prior probabilitas

antara atribut klasifikasi *case origin* dan *unit to charge* charge menggunakan software Weka rute Jakarta menuju Singapore.

```
Relation:     DATA MINING CGKSIN
Instances:    291
Attributes:   5
              Case Origin
              Type
              Ringkasan Comment
              Clustering Service
              Clustering Divisi
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier
```

Attribute	Class Customer Care (0.87)	Branch Office (0.1)	Customer Relation (0.02)	Digital Business (0.01)
Case Origin				
Call Center/Phone	3.0	2.0	2.0	1.0
Mail	1.0	2.0	1.0	1.0
Suggestion Form	237.0	26.0	2.0	2.0
Walk In	2.0	3.0	1.0	1.0
Web	13.0	1.0	3.0	1.0
Customer Care Email	6.0	1.0	1.0	1.0
Garuda Miles Email	1.0	1.0	1.0	2.0
[total]	263.0	36.0	11.0	9.0

Gambar 4.25 Perhitungan Prior Probabilitas Atribut Klasifikasi Case Origin dan Unit to Charge

4.10 Metode Bayes

Klasifikasi Bayes didasarkan pada Teorema Bayes. Pengklasifikasi Bayes adalah pengklasifikasi statistik. Pengklasifikasi Bayesian dapat memprediksi probabilitas keanggotaan kelas seperti probabilitas bahwa tupel tertentu milik kelas tertentu.

Teorema Bayes diambil dari nama Thomas Bayes. Ada dua jenis probabilitas

- Probabilitas Posterior [P(H/X)]
- Probabilitas Sebelumnya [P(H)]

di mana X adalah tupel data dan H adalah beberapa hipotesis.

Menurut Teorema Bayes,

P(H/X)= P(X/H)P(H) / P(X)……..(1)

Bayesian Belief Networks

Bayesian Belief Networks menentukan distribusi probabilitas bersyarat bersama. Mereka juga dikenal sebagai Jaringan Kepercayaan, Jaringan Bayesian, atau Jaringan Probabilistik.

- Jaringan Keyakinan memungkinkan independensi bersyarat kelas didefinisikan di antara subset variabel.
- Ini memberikan model grafis hubungan sebab akibat yang pembelajaran dapat dilakukan.
- Kita dapat menggunakan Jaringan Bayesian terlatih untuk klasifikasi.

Ada dua komponen yang mendefinisikan Jaringan Kepercayaan Bayesian

- Graf asiklik terarah
- Satu set tabel probabilitas bersyarat

Grafik asiklik terarah

- Setiap node dalam grafik asiklik berarah mewakili variabel acak.
- Variabel ini mungkin bernilai diskrit atau kontinu.
- Variabel-variabel ini mungkin sesuai dengan atribut aktual yang diberikan dalam data.

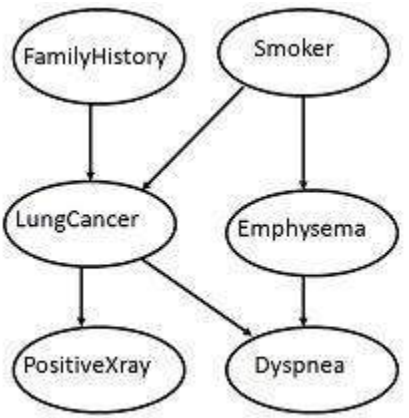
Directed Acyclic Graph

Diagram berikut menunjukkan grafik asiklik berarah untuk enam variabel Boolean.

Busur dalam diagram memungkinkan representasi pengetahuan kausal. Misalnya, kanker paru-paru dipengaruhi oleh riwayat keluarga seseorang yang mengidap kanker paru-paru, serta apakah orang tersebut perokok atau tidak. Perlu dicatat bahwa variabel PositiveXray tidak tergantung pada apakah pasien memiliki riwayat keluarga dengan kanker paru-paru atau bahwa pasien adalah perokok, mengingat bahwa kita tahu pasien menderita kanker paru-paru.

Tabel Probabilitas Bersyarat

Tabel probabilitas bersyarat untuk nilai variabel LungCancer (LC) yang menunjukkan setiap kemungkinan kombinasi nilai node induknya, FamilyHistory (FH), dan Smoker (S) adalah sebagai berikut:



4.11. Studi Kasus Metode Bayes

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang akan diolah. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data. Pre-Processing data berfokus pada sumber data untuk diolah dengan metode bayes. Data harus disiapkan untuk membuat target penentuan apakah nasabah puas atau tidak. Klasifikasi data dapat dilihat pada tabel 4.25, tabel 4.26, tabel 4.27, tabel 4.28, dan tabel 4.29. Kemudian untuk pengamatan inspeksi produk dan pelayanan dapat dilihat pada tabel 4.30.

a) Jenis Kelamin

Tabel 4.25 Klasifikasi Jenis Kelamin Nasabah

Jenis Kelamin	Count	Penjelasan
Laki-laki	59	Jumlah sampel nasabah Agency berjenis kelamin laki-laki yang mengisi kuesioner
Wanita	41	Jumlah sampel nasabah Agency berjenis kelamin wanita yang mengisi kuesioner

Jenis kelamin nasabah diklasifikasi menjadi 2 yaitu laki-laki dan wanita. Klasifikasi jenis kelamin dibuat agar dapat diketahui sampel data jenis kelamin nasabah yang menjadi objek penelitian.

b) Umur

Tabel 4.26 Klasifikasi Umur Nasabah

Umur	Count	Penjelasan
<25 tahun	13	Nasabah yang mengisi pilihan <25 tahun pada kuesioner tergolong pada kategori remaja
25-35 tahun	27	Nasabah yang mengisi pilihan 25-35 tahun pada kuesioner tergolong pada kategori dewasa
36-45 tahun	36	Nasabah yang mengisi pilihan 36-45 tahun pada kuesioner tergolong pada kategori usia pertengahan
46-55 tahun	13	Nasabah yang mengisi pilihan 46-55 tahun pada kuesioner tergolong pada kategori tua
>55 tahun	11	Nasabah yang mengisi pilihan >55 tahun pada kuesioner tergolong pada kategori lanjut usia

Umur nasabah diklasifikasi menjadi 5 klasifikasi remaja, dewasa, usia pertengahan, tua dan lanjut tua. Klasifikasi umur dibuat agar dapat diketahui data nasabah yang menjadi sampel objek penelitian.

c) Pekerjaan

Tabel 4.27 Klasifikasi Pekerjaan Nasabah

Pekerjaan	Count	Penjelasan
Mahasiswa	9	Jumlah nasabah Agency yang masih berstatus sebagai mahasiswa/ mahasiswi
Pegawai Negeri Sipil/BUMN/BUMD	14	Jumlah nasabah Agency yang bekerja sebagai PNS/BUMN/ BUMD
Pegawai Swasta dan Profesional	38	Jumlah nasabah Agency yang bekerja sebagai pegawai swasta dan profesional
Pengusaha	34	Jumlah nasabah Agency yang bekerja sebagai pengusaha

Pekerjaan	Count	Penjelasan
Lain-lain	5	Jumlah nasabah Agency yang tidak bekerja atau pekerjaannya tidak terdapat pada pilihan diatas

Pekerjaan nasabah diklasifikasi menjadi 5 klasifikasi mahasiswa, Pegawai Negeri Sipil/BUMN/BUMD, Pegawai Swasta & Profesional, Pengusaha, dan Lain-lain. Klasifikasi pekerjaan dibuat agar dapat diketahui pekerjaan nasabah yang menjadi sampel objek penelitian.

d) Lama berlangganan

Tabel 4.28 Klasifikasi Lama Berlangganan Nasabah

Lama Berlangganan	Count	Penjelasan
<1 tahun	11	Sudah menjadi nasabah dari Agency selama kurang dari satu tahun, tergolong baru
1-2 tahun	14	Sudah menjadi nasabah dari Agency diantara satu sampai dua tahun, cukup baru
2-5 tahun	39	Sudah menjadi nasabah dari Agency diantara dua sampai 5 tahun, tergolong cukup lama
>5 tahun	36	Sudah menjadi nasabah dari Agency lebih dari 5 tahun, tergolong loyal

Lama berlangganan nasabah diklasifikasi menjadi 4 klasifikasi yaitu baru, cukup baru, lama, dan loyal. Klasifikasi lama berlangganan dibuat agar dapat diketahui sudah berapa lama nasabah yang menjadi sampel menjadi nasabah Agency.

e) Klasifikasi

Tabel 4.29 Klasifikasi Kepuasan Nasabah

Klasifikasi	Count	Penjelasan
Sangat Memuaskan	17	Nasabah yang secara keseluruhan sangat puas dengan produk pelayanan asuransi, dengan mengisi nilai 5 pada kuesioner yang dibagikan
Memuaskan	56	Nasabah yang secara keseluruhan puas dengan produk pelayanan asuransi, dengan mengisi nilai 4 pada kuesioner yang dibagikan
Cukup Memuaskan	26	Nasabah yang secara keseluruhan puas dengan produk pelayanan asuransi, dengan mengisi nilai 3 pada kuesioner yang dibagikan
Tidak Memuaskan	1	Nasabah yang secara keseluruhan puas dengan produk pelayanan asuransi, dengan mengisi nilai 2 pada kuesioner yang dibagikan
Sangat Tidak Memuaskan	0	Nasabah yang secara keseluruhan sangat tidak puas dengan produk pelayanan asuransi, dengan mengisi nilai 1 pada kuesioner yang dibagikan

Kepuasan nasabah secara keseluruhan diklasifikasi menjadi 5 klasifikasi yaitu sangat tidak memuaskan, tidak memuaskan, cukup memuaskan, memuaskan, dan sangat memuaskan. Klasifikasi ditentukan dari hasil jawaban kuesioner.

f) Data Pengamatan

Tabel 4.30 Data Pengamatan Produk dan Pelayanan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
1	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
2	Wanita	25-35 tahun	Pengusaha	>5 tahun	Sangat Memuaskan
3	Laki-laki	<25 tahun	Pengusaha	>5 tahun	Sangat Memuaskan
4	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
5	Laki-laki	46-55 tahun	Pengusaha	>5 tahun	Memuaskan
6	Laki-laki	46-55 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
7	Wanita	36-45 tahun	Pengusaha	>5 tahun	Memuaskan
8	Laki-laki	46-55 tahun	Pengusaha	>5 tahun	Memuaskan
9	Laki-laki	46-55 tahun	Pegawai Negeri Sipil/BUMN/ BUMD	>5 tahun	Memuaskan
10	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
11	Wanita	25-35 tahun	Pengusaha	2-5 tahun	Memuaskan
12	Laki-laki	<25 tahun	Agen asuransi	2-5 tahun	Sangat Memuaskan
13	Laki-laki	25-35 tahun	Agen asuransi	>5 tahun	Sangat Memuaskan
14	Laki-laki	<25 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
15	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
16	Wanita	25-35 tahun	Pegawai Negeri Sipil/BUMN/ BUMD	2-5 tahun	Sangat Memuaskan
17	Laki-laki	46-55 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
18	Wanita	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
19	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	>5 tahun	Cukup Memuaskan
20	Wanita	36-45 tahun	Pengusaha	2-5 tahun	Memuaskan
21	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Sangat Memuaskan
22	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
23	Laki-laki	36-45 tahun	Agen asuransi	>5 tahun	Memuaskan
24	Laki-laki	46-55 tahun	Pengusaha	2-5 tahun	Sangat Memuaskan
25	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Tidak Memuaskan
26	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
27	Laki-laki	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
28	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Cukup Memuaskan
29	Laki-laki	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Memuaskan
30	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
31	Laki-laki	<25 tahun	Mahasiswa	<1 tahun	Cukup Memuaskan
32	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
33	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Cukup Memuaskan
34	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
35	Laki-laki	36-45 tahun	Pegawai Negeri Sipil/BUMN/ BUMD	2-5 tahun	Cukup Memuaskan
36	Wanita	46-55 tahun	Ibu rumah tangga	>5 tahun	Memuaskan
37	Laki-laki	46-55 tahun	Pegawai Negeri Sipil/BUMN/ BUMD	>5 tahun	Memuaskan
38	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	>5 tahun	Sangat Memuaskan
39	Wanita	25-35 tahun	Pegawai Negeri Sipil/BUMN/ BUMD	2-5 tahun	Memuaskan

Tabel 4.30 Data Pengamatan Produk dan Pelayanan (Lanjutan)

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
40	Wanita	36-45 tahun	Pengusaha	>5 tahun	Memuaskan
41	Laki-laki	36-45 tahun	Pengusaha	2-5 tahun	Cukup Memuaskan
42	Laki-laki	46-55 tahun	Pengusaha	2-5 tahun	Cukup Memuaskan
43	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Cukup Memuaskan
44	Laki-laki	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	2-5 tahun	Cukup Memuaskan
45	Laki-laki	<25 tahun	Mahasiswa	<1 tahun	Sangat Memuaskan
46	Wanita	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
47	Wanita	36-45 tahun	Pengusaha	>5 tahun	Memuaskan
48	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Cukup Memuaskan
49	Laki-laki	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Cukup Memuaskan
50	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
51	Laki-laki	46-55 tahun	Pengusaha	>5 tahun	Memuaskan
52	Laki-laki	36-45 tahun	Pengusaha	>5 tahun	Memuaskan
53	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
54	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
55	Laki-laki	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Memuaskan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
56	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Sangat Memuaskan
57	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Sangat Memuaskan
58	Laki-laki	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Sangat Memuaskan
59	Laki-laki	46-55 tahun	Pengusaha	2-5 tahun	Sangat Memuaskan
60	Laki-laki	<25 tahun	Pengusaha	1-2 tahun	Memuaskan
61	Laki-laki	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	>5 tahun	Cukup Memuaskan
62	Laki-laki	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	>5 tahun	Cukup Memuaskan
63	Laki-laki	25-35 tahun	Pengusaha	1-2 tahun	Memuaskan
64	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Sangat Memuaskan
65	Wanita	36-45 tahun	Pengusaha	2-5 tahun	Cukup Memuaskan
66	Wanita	>55 tahun	Pengusaha	>5 tahun	Cukup Memuaskan
67	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
68	Wanita	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
69	Wanita	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	2-5 tahun	Memuaskan
70	Laki-laki	25-35 tahun	Pengusaha	2-5 tahun	Cukup Memuaskan
71	Wanita	25-35 tahun	Pegawai Negeri Sipil/BUMN/BUMD	1-2 tahun	Memuaskan
72	Wanita	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	2-5 tahun	Memuaskan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
73	Wanita	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
74	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
75	Laki-laki	25-35 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Cukup Memuaskan
76	Wanita	>55 tahun	Pengusaha	>5 tahun	Memuaskan
77	Wanita	<25 tahun	Mahasiswa	<1 tahun	Cukup Memuaskan
78	Wanita	>55 tahun	Ibu rumah tangga	>5 tahun	Cukup Memuaskan
79	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Memuaskan
80	Laki-laki	46-55 tahun	Pengusaha	>5 tahun	Memuaskan
81	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Cukup Memuaskan
82	Wanita	25-35 tahun	Pegawai Negeri Sipil/BUMN/BUMD	1-2 tahun	Memuaskan
83	Wanita	>55 tahun	Pengusaha	1-2 tahun	Cukup Memuaskan
84	Laki-laki	>55 tahun	Pengusaha	2-5 tahun	Memuaskan

Responden	Jenis Kelamin	Umur Anda Saat Ini	Pekerjaan Anda Saat Ini	Lama Menjadi Nasabah Prudential	Klasifikasi
85	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Cukup Memuaskan
86	Laki-laki	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	2-5 tahun	Memuaskan
87	Laki-laki	25-35 tahun	Pengusaha	1-2 tahun	Sangat Memuaskan
88	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
89	Laki-laki	46-55 tahun	Pegawai Swasta dan Profesional	>5 tahun	Memuaskan
90	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
91	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Memuaskan
92	Wanita	25-35 tahun	Pegawai Swasta dan Profesional	1-2 tahun	Cukup Memuaskan
93	Wanita	36-45 tahun	Pegawai Negeri Sipil/BUMN/BUMD	>5 tahun	Cukup Memuaskan
94	Laki-laki	>55 tahun	Pengusaha	>5 tahun	Cukup Memuaskan
95	Laki-laki	<25 tahun	Mahasiswa	<1 tahun	Memuaskan
96	Wanita	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Memuaskan
97	Wanita	36-45 tahun	Pengusaha	2-5 tahun	Memuaskan
98	Laki-laki	25-35 tahun	Pengusaha	<1 tahun	Sangat Memuaskan
99	Laki-laki	25-35 tahun	Pengusaha	<1 tahun	Sangat Memuaskan
100	Laki-laki	36-45 tahun	Pegawai Swasta dan Profesional	2-5 tahun	Cukup Memuaskan

Berdasarkan data pengamatan, selanjutnya klasifikasi dengan metode *data mining* dilakukan dengan metode bayes. Dilakukan perhitungan manual dan juga dengan alat bantu yaitu *software* WEKA 3-8. Penggunaan *software* ini digunakan untuk melihat perbandingan apakah kesimpulan pada perhitungan manual dan menggunakan *software* memiliki hasil yang sama.

Perhitungan Metode Bayes

Data–data yang digunakan pada pengolahan data *mining* menggunakan data hasil kuesioner tingkat kepuasan pada produk dan pelayanan asuransi yang telah mengalami pengolahan berupa perubahan kode tingkat kepuasan menjadi tingkat kepuasan.

Tahap awal dalam perhitungan manual data mining menggunakan metode bayes adalah dengan cara menentukan probabilitas posterior klasifikasi terlebih dahulu. Probabilitas prior didapat dengan cara pembagian jumlah salah satu klasifikasi dengan jumlah koresponden, atau dapat dilihat pada rumus 4.1.

$$\frac{x_i}{\sum_{i=1}^{n} x} \quad \text{.....(4.1)}$$

Pada produk dan pelayanan asuransi memiliki probabilitas posterior yang memiliki nilai besar pada klasifikasi memuaskan dengan nilai diatas 0,5. Data hasil probabilitas posterior selengkapnya dapat dilihat pada tabel 4..31.

Tabel 4.31 Probabilitas Prior Produk dan Pelayanan Asuransi

Klasifikasi	Count	Probablitas Prior
Sangat Memuaskan	17	0,17
Memuaskan	56	0,56
Cukup Memuaskan	26	0,26
Tidak Memuaskan	1	0,01
Sangat Tidak Memuaskan	0	0,00

Tahap selanjutnya setelah mendapatkan probabilitas prior adalah mendapatkan probabilitas interaksi antara tiap atribut dengan klasifikasi. Probabilitas interaksi atribut dengan klasifikasi dapat dilihat pada tabel berikutnya. Terlihat pada tabel 4.32 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat Ketepatan solusi atau produk berada tingkat kepuasan sangat

puas dan klasifikasi sangat puas, memiliki probabilitas 0,31 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.32 Probabilitas Interaksi Ketepatan solusi atau produk

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Ketepatan solusi atau produk	Sangat Memuaskan	0,310344828	0,482758621	0,206896552			1
	Memuaskan	0,137931034	0,534482759	0,327586207			1
	Cukup Memuaskan		0,846153846	0,076923077	0,076923077		1
	Tidak Memuaskan						0
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.33 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kecepatan & kesigapan agen dalam memberikan respon/tanggapan berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,26 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.33 Probabilitas Interaksi Kecepatan & Kesigapan Agen dalam memberikan respon/tanggapan

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kecepatan & Kesigapan Agen dalam memberikan respon/ tanggapan	Sangat Memuaskan	0,260869565	0,608695652	0,130434783			1
	Memuaskan	0,12	0,62	0,26			1
	Cukup Memuaskan	0,25	0,4375	0,25	0,0625		1
	Tidak Memuaskan	0,1	0,3	0,6			1
	Sangat Tidak Memuaskan		1				1

Terlihat pada tabel 4.34 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kejelasan estimasi waktu dari agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,4 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.34 Probabilitas Interaksi Kejelasan estimasi waktu dari Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kejelasan estimasi waktu dari Agen	Sangat Memuaskan	0,4	0,466666667	0,133333333			1
	Memuaskan	0,173913043	0,586956522	0,239130435			1
	Cukup Memuaskan	0,0625	0,5625	0,34375	0,03125		1
	Tidak Memuaskan	0,142857143	0,571428571	0,285714286			1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.35 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat bahasa yang digunakan agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,17 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.35 Probabilitas Interaksi Bahasa yang digunakan Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Bahasa yang digunakan Agen	Sangat Memuaskan	0,166666667	0,5	0,333333333			1
	Memuaskan	0,214285714	0,535714286	0,232142857	0,017857143		1
	Cukup Memuaskan	0,052631579	0,684210526	0,263157895			1
	Tidak Memuaskan		1				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.36 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat keramahan agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,29 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel di bawah.

Tabel 4.36 Probabilitas Interaksi Keramahan Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Keramahan Agen	Sangat Memuaskan	0,294117647	0,529411765	0,176470588			1
	Memuaskan	0,1	0,58	0,3	0,02		1
	Cukup Memuaskan	0,071428571	0,571428571	0,357142857			1
	Tidak Memuaskan	0,5	0,5				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.37 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kemudahan agen untuk dihubungi berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,33 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.37 Probabilitas Interaksi Kemudahan Agen untuk dihubungi

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kemudahan Agen untuk dihubungi	Sangat Memuaskan	0,333333333	0,583333333	0,083333333			1
	Memuaskan	0,229166667	0,583333333	0,166666667	0,020833333		1
	Cukup Memuaskan	0,033333333	0,533333333	0,433333333			1
	Tidak Memuaskan	0,1	0,5	0,4			1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.38 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kemampuan mendengarkan agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,23 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.38 Probabilitas Interaksi Kemampuan mendengarkan agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kemampuan mendengar-kan agen	Sangat Memuaskan	0,235294118	0,588235294	0,176470588			1
	Memuaskan	0,166666667	0,592592593	0,222222222	0,018518519		1
	Cukup Memuaskan	0,115384615	0,461538462	0,423076923			1
	Tidak Memuaskan	0,333333333	0,666666667				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.39 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat ketersediaan waktu agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,44 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.39 Probabilitas Interaksi Ketersediaan waktu agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Ketersediaan waktu agen	Sangat Memuaskan	0,444444444	0,277777778	0,277777778			1
	Memuaskan	0,16	0,7	0,12	0,02		1
	Cukup Memuaskan		0,592592593	0,407407407			1
	Tidak Memuaskan	0,2		0,8			1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.40 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat keamanan & kenyamanan dalam melakukan transaksi berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,3 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.40 Probabilitas Interaksi Keamanan & Kenyamanan dalam Melakukan Transaksi

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Keamanan & kenya-manan dalam melakukan transaksi	Sangat Memuaskan	0,3	0,45	0,25			1
	Memuaskan	0,15625	0,59375	0,25			1
	Cukup Memuaskan	0,0625	0,5625	0,3125	0,0625		1
	Tidak Memuaskan						0
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.41 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kemampuan agen menyelesaikan permintaan & permasalahan berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,35 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.41 Probabilitas Interaksi Kemampuan Agen Menyelesaikan Permintaan & Permasalahan

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kemampuan agen menyelesaikan permintaan & permasalahan	Sangat Memuaskan	0,352941176	0,470588235	0,176470588			1
	Memuaskan	0,118644068	0,610169492	0,271186441			1
	Cukup Memuaskan	0,15	0,45	0,35	0,05		1
	Tidak Memuaskan	0,25	0,75				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.42 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat pengetahuan & keterampilan agen berada pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,22 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.42 Probabilitas Interaksi Pengetahuan & keterampilan Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Pengetahuan & keterampilan Agen	Sangat Memuaskan	0,217391304	0,47826087	0,304347826			1
	Memuaskan	0,20754717	0,509433962	0,264150943	0,018867925		1
	Cukup Memuaskan	0,047619048	0,714285714	0,238095238			1
	Tidak Memuaskan		1				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.43 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kemampuan memahami agen pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,28 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.43 Probabilitas Interaksi Kemampuan Memahami Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kemampuan memahami agen	Sangat Memuaskan	0,285714286	0,571428571	0,142857143			1
	Memuaskan	0,155172414	0,517241379	0,310344828	0,017241379		1
	Cukup Memuaskan	0,043478261	0,695652174	0,260869565			1
	Tidak Memuaskan	0,6	0,4				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.44 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kelengkapan agen pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,25 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel di bawah.

Tabel 4.44 Probabilitas Interaksi Kelengkapan Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kelengkapan agen	Sangat Memuaskan	0,25	0,45	0,3			1
	Memuaskan	0,166666667	0,592592593	0,222222222	0,018518519		1
	Cukup Memuaskan	0,142857143	0,523809524	0,333333333			1
	Tidak Memuaskan		0,8	0,2			1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.45 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat penampilan & pembawaan agen pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,36 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.45 Probabilitas Interaksi Penampilan & Pembawaan Agen

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Penampilan & pembawaan Agen	Sangat Memuaskan	0,363636364	0,363636364	0,272727273			1
	Memuaskan	0,116666667	0,6	0,266666667	0,016666667		1
	Cukup Memuaskan	0,0625	0,6875	0,25			1
	Tidak Memuaskan	0,5	0,5				1
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.46 probabilitas dari interaksi antara atribut dengan klasifikasi. Pada saat kepercayaan nasabah pada tingkat kepuasan sangat puas dan klasifikasi sangat puas, memiliki probabilitas 0,29 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.46 Probabilitas Interaksi Kepercayaan Nasabah

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Kepercayaan Nasabah	Sangat Memuaskan	0,294117647	0,529411765	0,176470588			1
	Memuaskan	0,15	0,6	0,25			1
	Cukup Memuaskan	0,130434783	0,47826087	0,347826087	0,043478261		1
	Tidak Memuaskan						0
	Sangat Tidak Memuaskan						0

Terlihat pada tabel 4.47 probabilitas dari interaksi antara jenis kelamin dengan klasifikasi. Pada saat jenis kelamin responden laki-laki dan klasifikasi sangat puas, memiliki probabilitas 0,25 pada tingkat kepuasan sangat puas. Probabilitas pada jenis kelamin dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 4.27 Probabilitas Interaksi Jenis Kelamin

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Jenis Kelamin	Laki-laki	0,254237288	0,491525424	0,254237288			1
	Wanita	0,048780488	0,658536585	0,268292683	0,024390244		1

Terlihat pada tabel 4.28 probabilitas dari interaksi antara usia dengan klasifikasi. Pada saat usia responden <25 tahun dan klasifikasi sangat puas, memiliki probabilitas 0,23 pada tingkat kepuasan sangat puas. Probabilitas pada usia dan klasifikasi lainnya dapat dilihat pada tabel dibawah.

Tabel 4.48 Probabilitas Interaksi Usia

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Usia	<25 tahun	0,230769231	0,615384615	0,153846154			1
	25-35 tahun	0,259259259	0,481481481	0,259259259			1
	36-45 tahun	0,138888889	0,583333333	0,25	0,027777778		1
	46-55 tahun	0,153846154	0,769230769	0,076923077			1
	>55 tahun		0,363636364	0,636363636			1

Terlihat pada tabel 4.49 probabilitas dari interaksi antara pekerjaan dengan klasifikasi. Pada saat pekerjaan responden mahasiswa dan klasifikasi sangat puas, memiliki probabilitas 0,11 pada tingkat kepuasan sangat puas. Probabilitas pada usia dan klasifikasi lainnya dapat dilihat pada tabel di bawah.

Tabel 4.49 Probabilitas Interaksi Pekerjaan

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Pekerjaan	Mahasiswa	0,111111111	0,666666667	0,222222222			1
	Pegawai Negeri Sipil/ BUMN/BUMD	0,071428571	0,571428571	0,357142857			1
	Pegawai Swasta dan Profesional	0,157894737	0,605263158	0,210526316	0,026315789		1
	Pengusaha	0,205882353	0,5	0,294117647			1
	Lain-lain	0,4	0,4	0,2			1

Terlihat pada tabel 4.50 probabilitas dari interaksi antara lama berlangganan dengan klasifikasi. Pada saat lama berlangganan responden <1 tahun dan klasifikasi sangat puas, memiliki probabilitas 0,27 pada tingkat kepuasan sangat puas. Probabilitas pada usia dan klasifikasi lainnya dapat dilihat pada tabel di bawah.

Tabel 4.50 Probabilitas Interaksi Lama Berlangganan

	Klasifikasi	Sangat Memuaskan	Memuaskan	Cukup Memuaskan	Tidak Memuasakan	Sangat Tidak Memuaskan	Total
Lama Berlangganan	<1 tahun	0,272727273	0,545454545	0,181818182			1
	1-2 tahun	0,285714286	0,428571429	0,285714286			1
	2-5 tahun	0,153846154	0,564102564	0,256410256	0,025641026		1
	>5 tahun	0,111111111	0,611111111	0,277777778			1

Setelah didapat probabilitas interaksi dari tiap atribut dengan klasifikasi, tahap selanjutnya dalam penggunaan klasifikasi dengan metode bayes adalah menghitung P (At|K) atau dikenal dengan probabilitas interaksi atribut dengan klasifikasi. P (At|K) didapat dengan cara pengalian tiap probabilitas interaksi dengan klasifikasi pada responden.

$$P(At|K) = P(At|K)_1 \times P(At|K)_2 \times \ldots \times P(At|K)_n \quad \text{.....(4.2)}$$

Setelah itu dilakukan perhitungan P(K) P(At|K). Perhitungan tersebut dengan cara pengalian dari probabilitas prior dari klasifikasi dengan probabilitas interaksi yang sudah didapat. Selanjutnya menghitung probabilitas posterior, probablitas posterior bisa didapat dengan cara sebagai berikut :

$$\frac{P(K).P(At|K)}{P(K)} = \quad \text{.....(4.3)}$$

Hasil dari perhitungan probabilitas posterior produk dan pelayanan asuransi Agency dengan menggunakan metode bayes dapat dilihat pada tabel 4.51.

Tabel 4.51 Hasil Perhitungan Probabilitas Posterior Produk dan Pelayanan Asuransi Agency

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
1	2,58138E-05	Memuaskan	0,56	1,44558E-05	0,44	3,2854E-05
2	2,81506E-12	Sangat Memuaskan	0,17	4,78561E-13	0,83	5,76579E-13
3	7,22629E-14	Sangat Memuaskan	0,17	1,22847E-14	0,83	1,48008E-14
4	7,78397E-05	Memuaskan	0,56	4,35902E-05	0,44	9,90687E-05
5	3,10775E-06	Memuaskan	0,56	1,74034E-06	0,44	3,95532E-06
6	2,89709E-06	Memuaskan	0,56	1,62237E-06	0,44	3,68721E-06
7	3,15748E-05	Memuaskan	0,56	1,76819E-05	0,44	4,01861E-05
8	2,50607E-06	Memuaskan	0,56	1,4034E-06	0,44	3,18954E-06
9	2,29036E-07	Memuaskan	0,56	1,2826E-07	0,44	2,91501E-07
10	2,3844E-06	Memuaskan	0,56	1,33527E-06	0,44	3,03469E-06
11	1,29052E-05	Memuaskan	0,56	7,2269E-06	0,44	1,64248E-05
12	2,08869E-12	Sangat Memuaskan	0,17	3,55077E-13	0,83	4,27803E-13
13	2,85051E-11	Sangat Memuaskan	0,17	4,84586E-12	0,83	5,83839E-12
14	2,01227E-05	Memuaskan	0,56	1,12687E-05	0,44	2,56108E-05
15	9,24185E-06	Memuaskan	0,56	5,17544E-06	0,44	1,17624E-05

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
16	3,7997E-16	Sangat Memuaskan	0,17	6,45949E-17	0,83	7,78252E-17
17	4,89462E-06	Memuaskan	0,56	2,74099E-06	0,44	6,22952E-06
18	6,24043E-05	Memuaskan	0,56	3,49464E-05	0,44	7,94236E-05
19	9,74044E-14	Cukup Memuaskan	0,26	2,53252E-14	0,74	3,42232E-14
20	1,05111E-05	Memuaskan	0,56	5,88622E-06	0,44	1,33778E-05
21	5,17433E-13	Sangat Memuaskan	0,17	8,79636E-14	0,83	1,0598E-13
22	2,82685E-05	Memuaskan	0,56	1,58303E-05	0,44	3,59781E-05
23	3,62393E-05	Memuaskan	0,56	2,0294E-05	0,44	4,61227E-05

Tabel 4.51 Hasil Perhitungan Probabilitas Posterior Produk dan Pelayanan Asuransi Agency (Lanjutan)

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
24	7,29205E-15	Sangat Memuaskan	0,17	1,23965E-15	0,83	1,49355E-15
25	2,58204E-30	Tidak Memuaskan	0,01	2,58204E-32	0,99	2,60812E-32
26	0,000116821	Memuaskan	0,56	6,54195E-05	0,44	0,000148681
27	1,06528E-05	Memuaskan	0,56	5,96557E-06	0,44	1,35581E-05
28	1,09348E-11	Cukup Memuaskan	0,26	2,84305E-12	0,74	3,84195E-12
29	7,61452E-06	Memuaskan	0,56	4,26413E-06	0,44	9,69121E-06
30	2,22678E-05	Memuaskan	0,56	1,247E-05	0,44	2,83409E-05
31	1,70228E-12	Cukup Memuaskan	0,26	4,42593E-13	0,74	5,98099E-13

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
32	1,17647E-05	Memuaskan	0,56	6,58824E-06	0,44	1,49733E-05
33	2,15391E-11	Cukup Memuaskan	0,26	5,60016E-12	0,74	7,56779E-12
34	2,73863E-05	Memuaskan	0,56	1,53363E-05	0,44	3,48553E-05
35	1,51056E-10	Cukup Memuaskan	0,26	3,92747E-11	0,74	5,30739E-11
36	7,3493E-06	Memuaskan	0,56	4,11561E-06	0,44	9,35366E-06
37	8,21988E-06	Memuaskan	0,56	4,60313E-06	0,44	1,04617E-05
38	1,27351E-16	Sangat Memuaskan	0,17	2,16497E-17	0,83	2,6084E-17
39	5,12488E-05	Memuaskan	0,56	2,86993E-05	0,44	6,52257E-05
40	0,000125011	Memuaskan	0,56	7,00063E-05	0,44	0,000159105
41	1,26194E-11	Cukup Memuaskan	0,26	3,28104E-12	0,74	4,43384E-12
42	1,03456E-11	Cukup Memuaskan	0,26	2,68986E-12	0,74	3,63494E-12
43	1,4268E-11	Cukup Memuaskan	0,26	3,70969E-12	0,74	5,01309E-12
44	3,90032E-11	Cukup Memuaskan	0,26	1,01408E-11	0,74	1,37038E-11
45	5,14859E-15	Sangat Memuaskan	0,17	8,75261E-16	0,83	1,05453E-15
46	3,73899E-05	Memuaskan	0,56	2,09383E-05	0,44	4,75871E-05
47	8,36789E-06	Memuaskan	0,56	4,68602E-06	0,44	1,065E-05
48	3,15057E-11	Cukup Memuaskan	0,26	8,19149E-12	0,74	1,10696E-11
49	7,98913E-11	Cukup Memuaskan	0,26	2,07717E-11	0,74	2,80699E-11
50	2,64185E-06	Memuaskan	0,56	1,47944E-06	0,44	3,36236E-06
51	2,54091E-06	Memuaskan	0,56	1,42291E-06	0,44	3,23388E-06
52	4,01266E-05	Memuaskan	0,56	2,24709E-05	0,44	5,10702E-05

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
53	4,96372E-05	Memuaskan	0,56	2,77968E-05	0,44	6,31746E-05
54	1,18802E-05	Memuaskan	0,56	6,65291E-06	0,44	1,51203E-05
55	2,68619E-05	Memuaskan	0,56	1,50426E-05	0,44	3,41878E-05
56	7,96621E-14	Sangat Memuaskan	0,17	1,35426E-14	0,83	1,63163E-14
57	7,96621E-14	Sangat Memuaskan	0,17	1,35426E-14	0,83	1,63163E-14
58	4,42295E-16	Sangat Memuaskan	0,17	7,51902E-17	0,83	9,05906E-17
59	7,28553E-15	Sangat Memuaskan	0,17	1,23854E-15	0,83	1,49222E-15
60	1,96561E-05	Memuaskan	0,56	1,10074E-05	0,44	2,50169E-05
61	3,87665E-10	Cukup Memuaskan	0,26	1,00793E-10	0,74	1,36207E-10
62	3,87665E-10	Cukup Memuaskan	0,26	1,00793E-10	0,74	1,36207E-10
63	7,97214E-06	Memuaskan	0,56	4,4644E-06	0,44	1,01464E-05
64	2,48969E-14	Sangat Memuaskan	0,17	4,23248E-15	0,83	5,09937E-15
65	1,16928E-10	Cukup Memuaskan	0,26	3,04013E-11	0,74	4,10828E-11
66	3,32428E-10	Cukup Memuaskan	0,26	8,64313E-11	0,74	1,16799E-10
67	2,30955E-05	Memuaskan	0,56	1,29335E-05	0,44	2,93942E-05
68	2,79841E-05	Memuaskan	0,56	1,56711E-05	0,44	3,56161E-05

Tabel 4.51 Hasil Perhitungan Probabilitas Posterior Produk dan Pelayanan Asuransi Agency (Lanjutan)

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
69	2,93096E-05	Memuaskan	0,56	1,64134E-05	0,44	3,73031E-05
70	7,73988E-12	Cukup Memuaskan	0,26	2,01237E-12	0,74	2,71942E-12
71	1,56054E-05	Memuaskan	0,56	8,73901E-06	0,44	1,98614E-05
72	1,85026E-05	Memuaskan	0,56	1,03615E-05	0,44	2,35488E-05
73	2,73984E-05	Memuaskan	0,56	1,53431E-05	0,44	3,48707E-05
74	3,95345E-06	Memuaskan	0,56	2,21393E-06	0,44	5,03166E-06
75	3,42886E-12	Cukup Memuaskan	0,26	8,91502E-13	0,74	1,20473E-12
76	1,12516E-05	Memuaskan	0,56	6,30088E-06	0,44	1,43202E-05
77	3,12756E-13	Cukup Memuaskan	0,26	8,13164E-14	0,74	1,09887E-13
78	3,34309E-11	Cukup Memuaskan	0,26	8,69203E-12	0,74	1,1746E-11
79	4,75809E-06	Memuaskan	0,56	2,66453E-06	0,44	6,05575E-06
80	2,28918E-06	Memuaskan	0,56	1,28194E-06	0,44	2,9135E-06
81	7,41839E-12	Cukup Memuaskan	0,26	1,92878E-12	0,74	2,60646E-12
82	2,31261E-05	Memuaskan	0,56	1,29506E-05	0,44	2,94332E-05
83	3,79337E-11	Cukup Memuaskan	0,26	9,86277E-12	0,74	1,33281E-11
84	1,16017E-05	Memuaskan	0,56	6,49697E-06	0,44	1,47658E-05
85	1,01437E-10	Cukup Memuaskan	0,26	2,63736E-11	0,74	3,564E-11
86	1,81239E-05	Memuaskan	0,56	1,01494E-05	0,44	2,30667E-05
87	8,0493E-14	Sangat Memuaskan	0,17	1,36838E-14	0,83	1,64865E-14
88	3,67463E-05	Memuaskan	0,56	2,05779E-05	0,44	4,67681E-05

Responden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
89	5,66541E-06	Memuaskan	0,56	3,17263E-06	0,44	7,21052E-06
90	4,18529E-05	Memuaskan	0,56	2,34376E-05	0,44	5,32673E-05
91	1,51809E-05	Memuaskan	0,56	8,50131E-06	0,44	1,93212E-05
92	7,3338E-11	Cukup Memuaskan	0,26	1,90679E-11	0,74	2,57674E-11
93	1,64072E-10	Cukup Memuaskan	0,26	4,26587E-11	0,74	5,76468E-11
94	1,21128E-10	Cukup Memuaskan	0,26	3,14932E-11	0,74	4,25583E-11
95	2,47471E-05	Memuaskan	0,56	1,38584E-05	0,44	3,14963E-05
96	4,18529E-05	Memuaskan	0,56	2,34376E-05	0,44	5,32673E-05
97	4,16652E-05	Memuaskan	0,56	2,33325E-05	0,44	5,30284E-05
98	7,48299E-14	Sangat Memuaskan	0,17	1,27211E-14	0,83	1,53266E-14
99	7,48299E-14	Sangat Memuaskan	0,17	1,27211E-14	0,83	1,53266E-14
100	6,57872E-12	Cukup Memuaskan	0,26	1,71047E-12	0,74	2,31144E-12

Berdasarkan hasil perhitungan diatas dapat dibuat rangkuman yang dapat dilihat pada tabel 4.52.

Tabel 4.52 Rangkuman Probabilitas Posterior Produk dan Pelayanan Asuransi Agency

Klasifikasi	Probablitas Prior	Probabilitas Posterior
Sangat Memuaskan	0,17	7,05266E-12
Memuaskan	0,56	0,001719532
Cukup Memuaskan	0,26	7,56975E-10
Tidak Memuaskan	0,01	2,60812E-32
Sangat Tidak Memuaskan	0,00	0

Probabilitas prior adalah probabilitas yang didapatkan dari data awal penelitian, sedangkan probabilitas posterior adalah Probabilitas yang telah diperbaiki setelah mendapat informasi tambahan atau setelah melakukan penelitian. Penarikan kesimpulan pada metode bayes menggunakan hasil data probabilitas posterior dengan cara melihat hasil dari probabilitas posterior. Kesimpulan diambil pada saat nilai probabilitas posterior tertinggi pada klasifikasi atau dapat di rumuskan sebagai berikut:

$h_{MAP} = \arg\max P(x|h)P(h)$ .....(4.4)

Pada penelitian ini produk dan pelayanan asuransi Agency berdasarkan data tabel 4.52 dapat ditarik kesimpulan bahwa tingkat kepuasan pelayanan yang diberikan perusahaan kepada responden berada pada tingkat kepuasan puas. Hasil dari ini memiliki kesamaan dengan hasil data mining dengan menggunakan software weka sehingga kesimpulan dapat diterima.

4.8.3 *Weka Data Mining*

Mula-mula, format tabel hasil pengamatan yang terdapat dalam *Microsoft Excell* diubah terlebih dahulu dari format .exe menjadi format .csv guna menginput data pada *software* WEKA 3-8. Tabel yang akan diubah ke dalam format .csv (*comma delimited*) dapat dilihat pada tabel 4.53 dibawah. Langkah-langkah yang diperlukan untuk mengubah format *file* menjadi .csv adalah sebagai berikut:

1. Buka *file excell* yang akan dikonversi ke .csv.
2. Klik *File*, kemudian *Save As.*
3. Pilih *folder* tempat menyimpan, kemudian pada '*Save As Type*' pilih 'CSV (*comma delimited*)'.
4. Klik '*Save*'.

Tabel 4.53 Data Percobaan Untuk *Data Mining*

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
4	4	4	4	3	4	4	3	4	3	3	3	3	4	3	Memuaskan
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	Sangat Memuaskan
5	4	4	4	5	4	4	5	4	5	4	4	5	4	4	Sangat Memuaskan
3	4	3	4	4	3	4	4	4	4	3	4	4	3	4	Memuaskan
5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	Memuaskan
4	5	4	3	3	2	3	4	4	5	3	4	3	4	5	Memuaskan
5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	Memuaskan
4	3	4	3	4	3	4	4	4	4	4	4	3	4	4	Memuaskan

Tabel 4.53 Data Percobaan Untuk *Data Mining* (Lanjutan)

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	Memuaskan
5	5	5	5	5	5	5	5	5	5	4	4	4	5	5	Memuaskan
4	4	4	4	5	5	5	4	4	4	4	4	4	5	5	Memuaskan
5	5	4	4	5	5	5	5	4	5	4	5	4	4	5	Sangat Memuaskan
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	Sangat Memuaskan
3	2	3	3	3	4	3	4	4	3	4	4	2	3	3	Memuaskan
5	5	5	5	5	5	5	5	4	4	4	4	4	4	4	Memuaskan
4	3	3	4	4	4	3	5	4	4	5	5	5	4	4	Sangat Memuaskan
4	4	4	4	3	3	4	3	4	4	4	3	4	3	4	Memuaskan
5	5	5	4	3	4	4	4	3	4	3	3	4	3	4	Memuaskan

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	Cukup Memuaskan
5	5	5	4	4	5	5	4	5	5	5	5	5	3	5	Memuaskan
5	5	4	4	5	5	5	5	5	5	4	4	4	4	5	Sangat Memuaskan
3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	Memuaskan
3	4	4	4	4	4	5	4	4	4	4	5	4	4	4	Memuaskan
5	4	5	4	3	2	3	2	5	4	5	4	3	2	3	Sangat Memuaskan
3	3	3	4	4	4	4	4	3	3	4	4	4	4	3	Tidak Memuaskan
3	4	4	3	4	4	4	3	4	4	3	3	4	3	4	Memuaskan
4	4	3	4	5	4	5	4	5	3	4	4	5	4	5	Memuaskan
5	4	4	5	5	4	4	5	4	3	4	4	4	4	4	Cukup Memuaskan
5	3	4	5	5	3	4	4	4	4	4	5	5	4	4	Memuaskan
4	3	3	4	4	4	4	4	4	4	4	3	5	4	4	Memuaskan
4	5	4	5	4	4	4	5	4	4	3	4	5	5	4	Cukup Memuaskan
4	4	4	4	4	3	3	3	4	4	4	4	5	4	3	Memuaskan
4	4	4	4	4	3	3	4	5	5	4	4	5	5	5	Cukup Memuaskan
4	4	3	4	4	3	4	4	4	4	4	3	4	4	3	Memuaskan
3	2	3	3	3	3	3	3	3	3	4	3	3	4	3	Cukup Memuaskan
4	4	3	3	4	4	2	4	4	4	4	3	2	4	3	Memuaskan
3	4	3	3	2	2	4	4	3	4	4	3	4	3	4	Memuaskan
4	3	4	4	2	3	3	4	3	2	3	2	3	3	3	Sangat Memuaskan
3	3	2	2	3	4	3	4	3	3	4	4	2	3	4	Memuaskan

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
3	1	2	4	4	2	2	3	3	4	2	3	3	4	3	Memuaskan
4	2	5	5	4	4	4	5	3	4	3	3	2	4	3	Cukup Memuaskan
4	4	3	3	4	2	3	4	4	3	4	4	3	3	4	Cukup Memuaskan
4	4	3	4	4	3	4	4	3	4	4	4	3	3	4	Cukup Memuaskan
4	4	4	3	5	2	3	2	4	4	4	5	5	5	4	Cukup Memuaskan
4	4	4	4	5	4	4	4	4	4	4	4	5	4	4	Sangat Memuaskan
4	4	4	5	5	4	4	4	4	4	4	4	4	4	4	Memuaskan
5	5	4	5	5	4	5	4	5	4	5	5	5	5	5	Memuaskan
5	4	4	3	5	3	5	3	5	4	5	4	5	4	4	Cukup Memuaskan
4	2	3	4	4	2	3	3	4	4	4	3	4	4	4	Cukup Memuaskan
5	5	4	5	5	5	4	5	5	5	5	5	5	5	4	Memuaskan
4	3	3	4	4	4	3	4	4	4	3	4	4	4	4	Memuaskan
4	5	3	3	4	2	3	3	3	2	2	3	3	2	4	Memuaskan
3	4	3	3	3	2	4	4	3	4	3	2	3	3	4	Memuaskan
3	2	3	3	4	3	3	4	4	3	5	4	4	4	3	Memuaskan

Tabel 4.53 Data Percobaan Untuk *Data Mining* (Lanjutan)

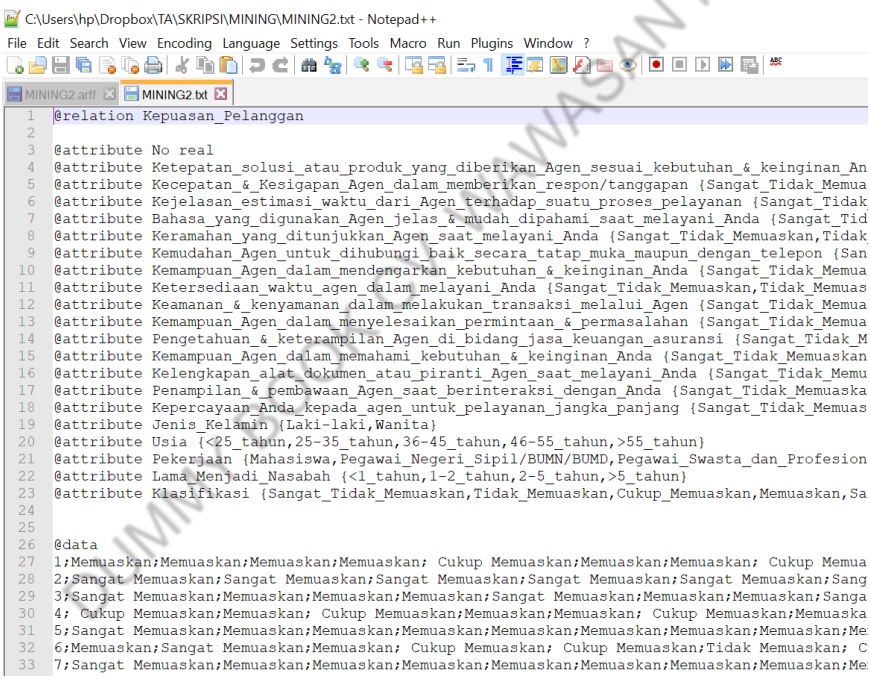
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
4	3	2	3	3	3	4	3	3	2	3	3	4	4	4	Memuaskan
4	5	5	4	4	4	4	4	4	3	4	2	4	5	4	Sangat Memuaskan
4	5	5	4	4	4	4	4	4	3	4	2	4	5	4	Sangat Memuaskan
4	2	3	3	5	4	4	4	4	4	5	4	4	4	4	Sangat Memuaskan
4	3	2	4	4	4	2	5	4	3	4	4	3	4	3	Sangat Memuaskan
4	5	4	4	4	4	4	4	4	4	5	5	4	4	4	Memuaskan
4	2	3	4	3	3	3	3	4	3	3	4	3	5	4	Cukup Memuaskan
4	2	3	4	3	3	3	3	4	3	3	4	3	5	4	Cukup Memuaskan
4	4	4	4	5	4	4	4	4	5	3	4	4	5	5	Memuaskan
4	3	4	4	5	4	4	5	4	5	4	3	4	5	4	Sangat Memuaskan
4	2	3	4	4	3	4	5	4	3	5	4	3	4	4	Cukup Memuaskan
4	4	4	4	4	3	3	2	4	4	4	4	4	4	3	Cukup Memuaskan
4	4	3	4	4	3	4	3	4	4	4	4	4	4	3	Memuaskan
4	5	4	4	5	4	4	4	5	4	5	4	4	4	4	Memuaskan
4	4	4	4	4	4	4	3	4	4	5	4	4	4	4	Memuaskan
5	4	3	5	5	4	4	3	5	4	5	4	4	4	4	Cukup Memuaskan
4	4	4	4	4	4	4	3	4	4	4	4	4	4	3	Memuaskan
4	4	4	4	4	3	4	4	5	5	5	4	4	4	4	Memuaskan
5	4	3	5	5	4	4	3	4	4	4	4	4	4	4	Memuaskan
5	5	5	4	5	5	5	5	5	5	5	5	4	4	5	Memuaskan

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
4	4	4	4	3	4	4	3	5	5	4	4	4	4	5	Cukup Memuaskan
4	4	4	5	5	4	4	4	4	4	4	4	4	5	4	Memuaskan
5	5	4	5	5	4	4	4	4	4	4	4	5	4	4	Cukup Memuaskan
4	4	3	4	4	4	4	3	4	4	4	4	4	4	3	Cukup Memuaskan
4	4	3	4	5	3	4	3	4	4	5	4	5	5	4	Memuaskan
5	4	5	5	4	4	5	4	4	4	4	4	4	4	4	Memuaskan
4	4	4	5	4	3	5	4	4	4	5	4	4	4	4	Cukup Memuaskan
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	Memuaskan
4	4	4	4	4	4	3	3	4	4	4	3	4	3	4	Cukup Memuaskan
4	4	4	4	5	4	4	3	4	4	4	4	4	4	4	Memuaskan
4	3	2	3	4	2	3	2	4	3	3	3	4	4	3	Cukup Memuaskan
4	2	3	4	4	4	4	3	4	4	3	2	2	3	4	Memuaskan
5	4	5	4	4	4	4	4	4	4	4	4	4	5	5	Sangat Memuaskan
4	5	4	4	4	4	3	3	4	3	2	3	4	4	4	Memuaskan
3	4	2	4	4	3	3	4	4	4	3	4	3	4	4	Memuaskan
4	4	3	3	4	3	3	4	4	3	3	3	3	4	4	Memuaskan
4	4	4	5	4	4	4	4	4	4	4	4	4	4	4	Memuaskan
5	3	3	5	4	3	3	3	4	4	5	4	4	4	3	Cukup Memuaskan
4	3	2	4	3	3	4	2	3	4	5	4	4	4	3	Cukup Memuaskan
4	4	3	4	4	3	4	3	4	4	4	4	4	4	3	Cukup Memuaskan
5	5	4	5	5	4	4	4	4	4	4	4	4	4	4	Memuaskan

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Klasifikasi
4	4	3	3	4	3	3	4	4	3	3	3	3	4	4	Memuaskan
4	4	4	4	5	3	4	4	3	2	3	3	3	4	3	Memuaskan
5	4	4	5	5	4	4	4	5	4	4	4	4	5	4	Sangat Memuaskan
5	4	4	5	5	4	4	4	5	4	4	4	4	5	4	Sangat Memuaskan
4	3	4	4	4	3	4	4	3	4	4	3	3	3	4	Cukup Memuaskan

Dimana Q1 sampai Q15 adalah atribut kepuasan pelanggan yang masing-masing penjelasannya dapat dilihat di lampiran atau pada tabel 4.53. Setelah mengubah tabel diatan kedalam bentuk .csv, langkah selanjutnya adalah menentukan attribute dalam tabel agar dapat dibaca oleh aplikasi *weka 3-8*. Langkah-langkah yang diperlukan untuk mengubah atribut agar dapat dibaca oleh aplikasi *weka* adalah sebagai berikut:

1. Klik kanan pada *file excell* yang telah dikonversi ke .csv.
2. Klik *edit with notepad ++*
3. Buat a*ttribute* seperti pada gambar 4.26 dibawah
4. Klik '*Save'*, tambahkan '.arff' dibelakang nama *file* agar dapat dibaca oleh aplikasi *weka*

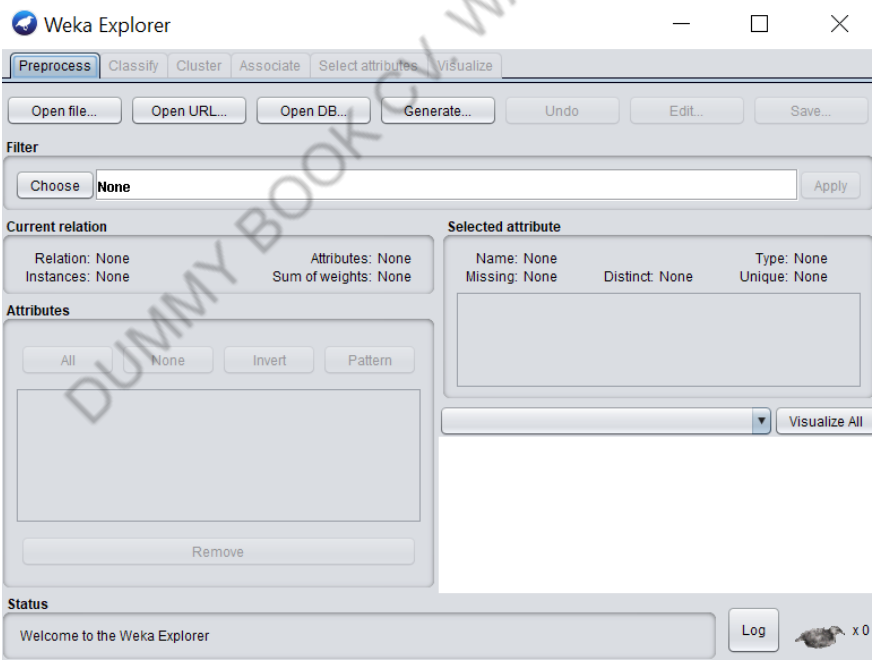


Gambar 4.26 Attribute *Weka Data Mining*

Kemudian, buka *software* WEKA 3-8. Tampilan awal dari *software* tersebut terdapat pada Gambar 4.27.

Gambar 4.27 Tampilan *Software* WEKA 3-8

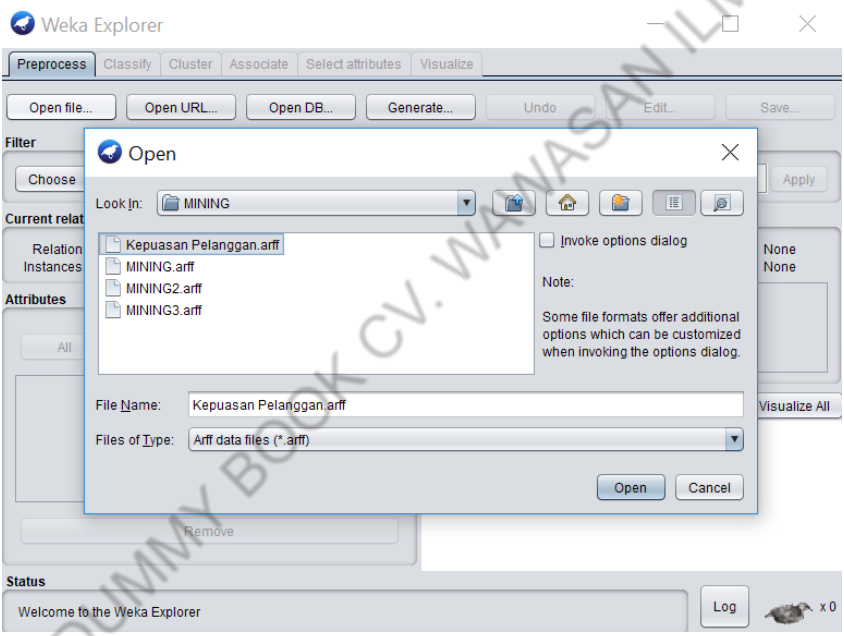
Pada tampilan *software* WEKA tersebut, terdapat 4 opsi utama dalam penggunaan *data mining*, yaitu *Explorer, Experimenter, KnowledgeFlow, Workbench* dan *Simple CLI*. Untuk membuat *bayes classifier*, pilihan yang digunakan adalah *Explorer*. Setelah opsi *Explorer* di klik, maka akan muncul tampilan seperti Gambar 4.28.



Gambar 4.28 Tampilan *Software Weka Explorer*

Halaman awal pada Gambar 4.7 tersebut langsung menunjukkan bagian *pre-processing*. *Pre-process* yang mana merupakan tahap awal dalam *data mining* ini bertujuan untuk meng-*input* data yang akan dilakukan perhitungan menggunakan metode *bayes*. Tahap ini nantinya juga akan menunjukkan faktor-faktor yang berpengaruh terhadap hasil dari objek yang sedang diteliti. Langkah-langkah untuk meng-*input file* pada *pre-processing* adalah sebagai berikut:

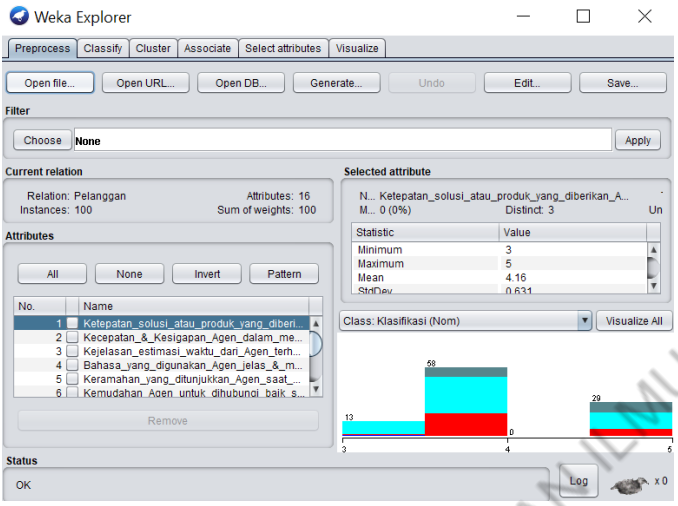
1. Klik opsi '*Open File*' pada bagian kiri atas.
2. Buka *folder* tempat *file* sebelumnya disimpan.
3. Pilih *file* yang akan di-*input*. Kotak Dialog yang ditampilkan dapat dilihat pada Gambar 4.29.



Gambar 4.29 *File* dengan *format* .arff

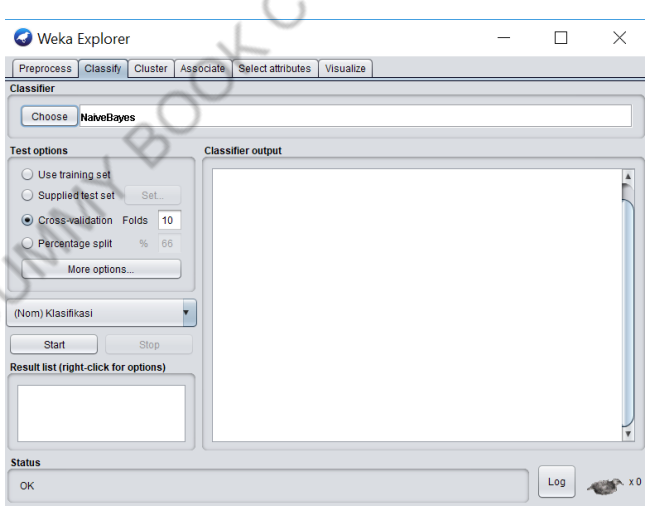
4. Klik '*Open*'

Setelah itu, apabila *file* telah dipilih maka *software* akan menampilkan informasi-informasi yang berada pada *file* tersebut, seperti faktor yang mempengaruhi variabel respon dan juga keputusannya. Tampilan tersebut dapat dilihat pada Gambar 4.30.



Gambar 4.30 *Input software Weka*

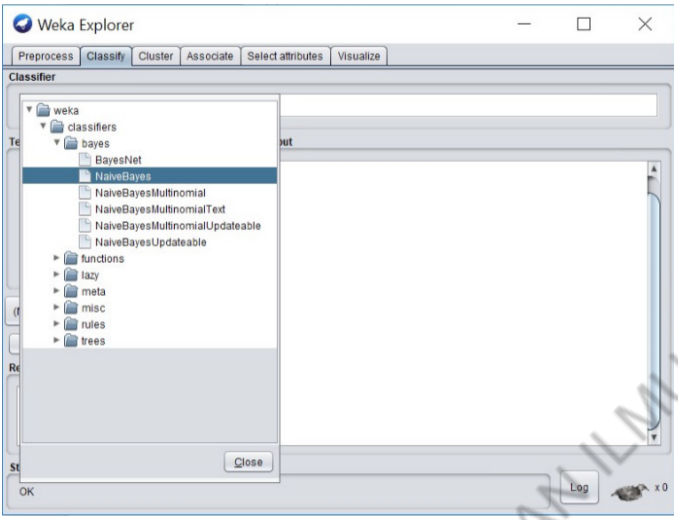
Pada tampilan di atas dapat dilihat atribut yang mempengaruhi kepuasan pelanggan, beserta nilai *maximum, minimum,* rata-rata, dan standar deviasi dari setiap atribut. Selanjutnya, pilih menu *'Classify'* di sebelah menu *'Preprocessing'*. Tampilan dari *'Classify'* dapat dilihat pada Gambar 4.31.



Gambar 4.31 Tampilan *Classify*

Langkah selanjutnya adalah sebagai berikut:

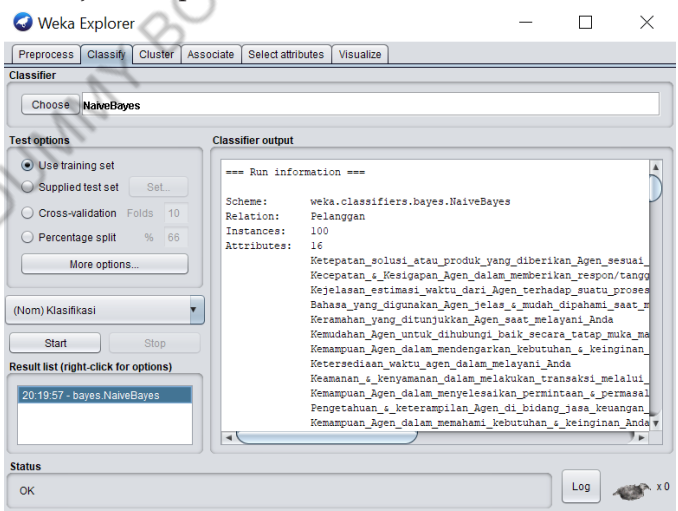
1. Klik *'Choose'*.
2. Pilih opsi *'Bayes'*. Maka akan muncul tampilan yang ditunjukkan pada Gambar 4.32.



Gambar 4.32 Opsi '*bayes*' pada *classify*

3. Pilih opsi '*NaiveBayes*'
4. Pada '*Test Options*', pilih '*Use Training Set*'.
5. Klik '*Start*.

Selanjutnya, maka *software* akan mengeluarkan *output* berdasarkan *input* yang diterima pada tahap *preprocessing*, dan juga pilihan yang dibuat pada tahap *classify*. Tampilan dari *output software* ditunjukkan pada Gambar 4.33.



Gambar 4.33 *Output Software Weka*

Hasil dari perhitungan data mining menggunakan *software weka* pada produk dan pelayanan Agency menghasilkan statistik seperti Gambar 4.34, pada gambar tersebut data yang digunakan memiliki tingkat keberhasilan data dalam mengklasifikasi hasil oleh *weka* berada pada 99%. Dengan nilai keberhasilan yang tinggi tersebut memiliki mean absolute error yang rendah yaitu berada pada nilai 0.0289, sehingga dapat dikatakan bahwa hasil tersebut memiliki margin error yang sangat kecil. Pada hasil statistik *coverage of case* dapat dilihat bahwa hasil berada diatas *level coverage of case* yang sudah ditentukan. Dengan melihat hasil statistik data mining menggunakan *software* weka dapat ditarik kesimpulan bahwa hasil data mining dapat diterima.

```
=== Summary ===

Correctly Classified Instances          99               99      %
Incorrectly Classified Instances         1                1      %
Kappa statistic                          0.98
Mean absolute error                      0.0289
Root mean squared error                  0.1193
Relative absolute error                  5.7763 %
Root relative squared error             23.8597 %
Coverage of cases (0.95 level)          99      %
Mean rel. region size (0.95 level)      53      %
Total Number of Instances              100
```

Gambar 4.34 Statistik *Weka Data Mining* Produk dan Pelayanan Asuransi

Hasil klasifikasi berdasarkan data mining menggunakan *software weka* pada produk dan pelayanan asuransi menghasilkan klasifikasi yang lebih spesifik jika dibandingkan dengan hasil perhitungan manual menggunakan metode bayes. Pada hasil tersebut terlihat jumlah dari tingkat kepuasan berdasarkan hubungan tingkat kepuasan yang ada. Namun hasil klasifikasi dari *data mining* ini sama dengan hasil dari perhitungan manual. Secara umum bahwa tingkat kepuasan pelayanan yang diberikan pada produk dan pelayanan asuransi berada pada posisi katagori memuaskan dengan jumlah koresponden sebanyak 51 orang yang berada pada kategori tersebut, sedangkan pada kategori sangat memuaskan sebanyak 25 koresponden. Hasil klasifikasi menggunakan *software weka* dapat dilihat pada gambar 4.35.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  0 |  a = Sangat_Tidak_Memuaskan
  0  1  0  0  0 |  b = Tidak_Memuaskan
  0  0 17  7  2 |  c = Cukup_Memuaskan
  0  1  4 41 10 |  d = Memuaskan
  0  0  1  3 13 |  e = Sangat_Memuaskan
```

Gambar 4.35 Matriks tingkat kepuasan Produk dan Pelayanan Asuransi

4.12. Latihan Soal

1. Data pengecekan pipa baja meliputi empat parameter pemeriksaan, yaitu penyerutan *inner* miring, penyerutan *inner* bergelombang, berlubang dan hasil las tidak kuat. Dilakukan pengambilan sampel cacat sebanyak 35 pipa yang diinspeksi berdasarkan keempat parameter kecacatan. Kemudian keempat parameter tersebut akan diklasifikasikan kedalam kelas *repair* dan *QC Pass*. Berikut dapat dilihat data pengamatan inspeksi yang telah dilakukan :

Miring	Bergelombang	Berlubang	Las Tidak Kuat	Class
No	No	Yes	Yes	Repair
Yes	Yes	No	No	Repair
No	Yes	No	Yes	Repair
Yes	No	Yes	No	Repair
Yes	No	No	No	QC Pass
No	Yes	No	No	QC Pass
No	No	Yes	Yes	QC Pass
No	No	No	Yes	Repair
No	No	No	No	QC Pass
Yes	Yes	Yes	Yes	Repair
No	No	Yes	No	Repair
Yes	No	Yes	No	Repair
Yes	Yes	No	No	Repair
No	No	No	Yes	Repair
Yes	No	No	No	QC Pass
No	Yes	No	No	QC Pass
No	No	Yes	Yes	Repair

Yes	Yes	No	No	Repair
No	Yes	No	Yes	Repair
Yes	No	Yes	No	Repair
Yes	No	No	No	QC Pass
No	Yes	No	No	QC Pass
No	No	Yes	Yes	QC Pass
No	No	Yes	No	Repair
No	No	No	No	QC Pass
Yes	Yes	Yes	Yes	Repair
No	No	Yes	No	Repair
Yes	No	Yes	No	Repair
Yes	Yes	No	No	Repair
No	No	Yes	No	Repair
Yes	No	No	No	QC Pass
No	Yes	No	No	QC Pass
No	No	Yes	Yes	Repair
Yes	Yes	No	No	Repair
No	Yes	No	Yes	Repair

a. Berapakah nilai Gini Index Keselutuhannya pada Tabel diatas ?

b. Berapakah nilai *Entropy* dari kelas Repair pada Tabel 1 diatas?

c. Buatlah analisanya dengan menggunakan software Weka!

2. Berikut adalah Heart disease binary data

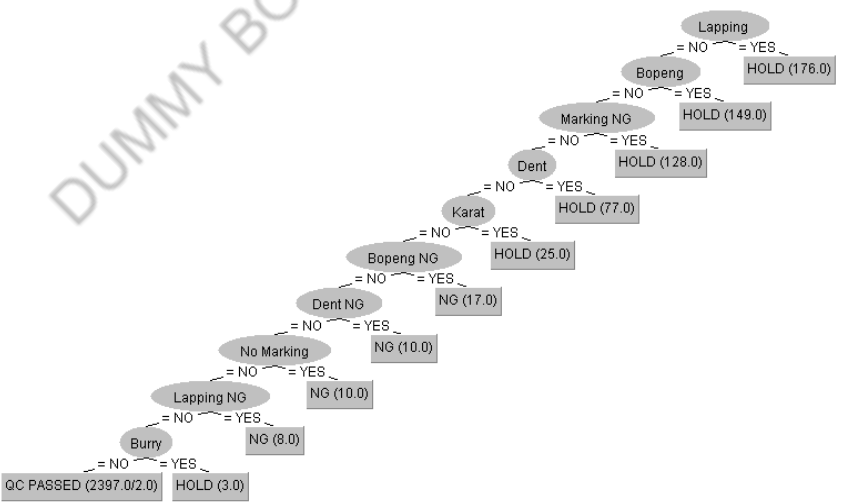
Sebuah tim peneliti mengumpulkan tim peneliti untuk mengumpulkan dan mempublikasikan informasi detail mengenai factor yang mempengaruhi heart disease. Variabel meliputi age, sex, cholesterol levels, maximum heart rate, and more.

Gunakan data ini untuk membuat decision tree menggunakan software dan hitungan rumus menggunakan Microsoft Excel.

Kolom	Deskripsi
Age	The age of the patient
Sex	The sex of the patient: Male or Female
Chest Pain Type	The chest pain type: 1, 2, 3, or 4
Rest Blood Pressure	The resting blood pressure of the patient, in mm Hg
Cholesterol	The serum cholesterol level of the patient, in mg/dl

Kolom	Deskripsi
Fasting Blood Sugar	Whether the patient fasting blood sugar > 120 mg/dl: True or False
Rest ECG	The resting electrocardiographic results: 0, 1, or 2
Max Heart Rate	The maximum heart rate achieved
Exercise Angina	Whether the patient has exercise-induced angina: Yes or No
Old Peak	The ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment: 1, 2, or 3
Major Vessels	The number of major vessels colored by fluoro-sopy: 0, 1, 2, 3, or 4
Thal	The type of defect: Normal, Fixed, or Reversible
Heart Disease	The binary response that indicates whether the patient has heart disease: Yes or No

3. Berikut adalah decision tree untuk menentukan apakah produk Good, Repair atau Dispose. Buatkan *if then rule* nya.



4. Sebuah perusahaan asuransi ingin melakukan analisis data tingkat kepuasan nasabah. Perusahaan ingin mengetahui segmentasi nasabah sehingga dapat dirumuskan strategi yang tepat untuk setiap segmen. Silahkan download https://docs.google.com/spreadsheets/d/1jPX4mVz_AzIHMVF_r18pBXWsQs8s3zv_/edit?usp=sharing&ouid=107032425374168372161&rtpof=true&sd=true dan akses https://rpubs.com/anikbibib/828643 sebagai petunjuk pengerjaan dan coding telah tersedia. Dalam contoh dilakukan pembagian data menjadi 80% (data training) dan 20% (data testing). Silahkan lakukan pengolahan data dengan pembagian 90% (training) dan 10% (testing) untuk pengujian model klasifikasi. Kerjakan dengan menggunakan software R.
   a. Bandingkan kedua hasil akurasi model!
   b. Apakah ada perbedaan hasil variable penting dalam klasifikasi?
   c. Berikan kesimpulan dari hasil pohon keputusan yang diperoleh!
   d. Berikan penjelasan singkat tentang packages rpart yang digunakan!

BAB V
DASAR DAN PENGEMBANGAN ANALISIS ASOSIASI

5.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Mahasiswa mampu menganalisis, menafsirkan data dan informasi dan membuat keputusan yang tepat berdasarkan pendekatan analisis asosiasi. Bab ini bertujuan untuk mengenalkan konsep analisis asosiasi.

Materi yang diberikan :

1. Konsep Dasar Analisis Asosiasi.
2. Association Rules.
3. Dasar Asosiasi

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah menganalisis, menafsirkan data dan informasi dan membuat keputusan yang tepat berdasarkan pendekatan analisis asosiasi. Capaian pembelajaran lulusan pada bab ini adalah capaian pembelajaran pengetahuan adalah mampu menguasai prinsip dan teknik perancangan sistem terintegrasi dengan pendekatan sistem.

5.2 Konsep Dasar Analisis Asosiasi

Pola yang sering menunjukkan hubungan yang menarik antara pasangan atribut-nilai yang sering terjadi dalam kumpulan data yang diberikan. Misalnya, kita mungkin menemukan bahwa pasangan atribut-nilai age = pemuda dan kredit = OK terjadi pada 20% tupel data yang menggambarkan pelanggan Elektronik yang membeli komputer. Kita dapat menganggap setiap pasangan atribut-nilai sebagai item, sehingga pencarian untuk pola frequent ini dikenal sebagai frequent pattern mining atau frequent itemset mining.

Selanjutnya akan melihat bagaimana aturan asosiasi diturunkan dari pola yang sering terjadi, di mana asosiasi biasanya digunakan untuk menganalisis pola pembelian pelanggan di sebuah toko. Analisis tersebut berguna dalam banyak proses pengambilan keputusan seperti penempatan produk, desain katalog, dan pemasaran silang.

Association rule adalah suatu prosedur yang mencari hubungan atau relasi antara satu *item* dengan *item* lainnya. *Association rule* biasanya menggunakan "*if*" dan "*then*" misalnya "*if* A *then* B *and* C", hal ini menunjukkan jika A maka B dan C. Dalam menentukan *association rule* perlu ditentukan *support* dan *confidence* untuk membatasi apakah *rule* tersebut *interesting* atau tidak.

Association rule berguna untuk menemukan hubungan penting antar *item* dalam setiap transaksi, hubungan tersebut dapat menandakan kuat tidaknya suatu aturan dalam asosiasi, Tujuan *association rule* adalah untuk menemukan keteraturan dalam data. *Association rule* dapat digunakan untuk mengidentifikasi *item-item* produk yang mungkin dibeli secara bersamaan dengan produk lain, atau dilihat secara bersamaan saat mencari informasi mengenai produk tertentu. Dalam pencarian *association rule*, diperlukan suatu variabel ukuran kepercayaan *(interestingness measure)* yang dapat ditentukan oleh *user*, untuk mengatur batasan sejauh mana dan sebanyak apa hasil *output* yang diinginkan oleh *user*.

Asosiasi berguna untuk mengungkap hubungan yang menarik yang tersembunyi dalam dataset besar.

Hubungan yang terungkap tersebut dapat direpresentasikan dalam bentuk aturan asosiasi (*association rules*) atau himpunan item yang sering muncul (*sets of frequent items*).

Frequent pattern: pola (himpunan item, sekuens, atau struktur) yang sering muncul dalam basis data [AIS93]

Dari data pembelian Ibu ibu di supermarket kita dapat melihat pola pembeliannya.

Tabel 5.1 Data Pembelian

TID	Item
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}

TID	Item
5	{Bread, Milk, Diapers, Cola}

Motivasinya mencari keteraturan (*regularities*) dalam data

Pola yang sering menunjukkan hubungan yang menarik antara pasangan atribut-nilai yang sering terjadi dalam kumpulan data yang diberikan. Misalnya, kita mungkin menemukan bahwa pasangan atribut- nilai age = pemuda dan kredit = OK terjadi pada 20% tupel data yang menggambarkan pelanggan AllElectronics yang membeli komputer. Kita dapat menganggap setiap pasangan atribut-nilai sebagai item, sehingga pencarian untuk pola frequent ini dikenal sebagai frequent pattern mining atau frequent itemset mining. Selanjutnya akan melihat bagaimana aturan asosiasi diturunkan dari pola yang sering terjadi, di mana asosiasi biasanya digunakan untuk menganalisis pola pembelian pelanggan di sebuah toko. Analisis tersebut berguna dalam banyak proses pengambilan keputusan seperti penempatan produk, desain katalog, dan pemasaran silang.

5.3 Association Rules

Association rule adalah suatu prosedur yang mencari hubungan atau relasi antara satu *item* dengan *item* lainnya. *Association rule* biasanya menggunakan "*if*" dan "*then*" misalnya "*if* A *then* B *and* C", hal ini menunjukkan jika A maka B dan C. Dalam menentukan *association rule* perlu ditentukan *support* dan *confidence* untuk membatasi apakah *rule* tersebut *interesting* atau tidak.

Association rule berguna untuk menemukan hubungan penting antar *item* dalam setiap transaksi, hubungan tersebut dapat menandakan kuat tidaknya suatu aturan dalam asosiasi, Tujuan *association rule* adalah untuk menemukan keteraturan dalam data. *Association rule* dapat digunakan untuk mengidentifikasi *item-item* produk yang mungkin dibeli secara bersamaan dengan produk lain, atau dilihat secara bersamaan saat mencari informasi mengenai produk tertentu. Dalam pencarian *association rule*, diperlukan suatu variabel ukuran kepercayaan *(interestingness measure)* yang dapat ditentukan oleh *user*, untuk mengatur batasan sejauh mana dan sebanyak apa hasil *output* yang diinginkan oleh *user*.

1. Dataset

Dataset : Data Supermaket Atribut :

Tabel 5.2 Dataset Supermarket

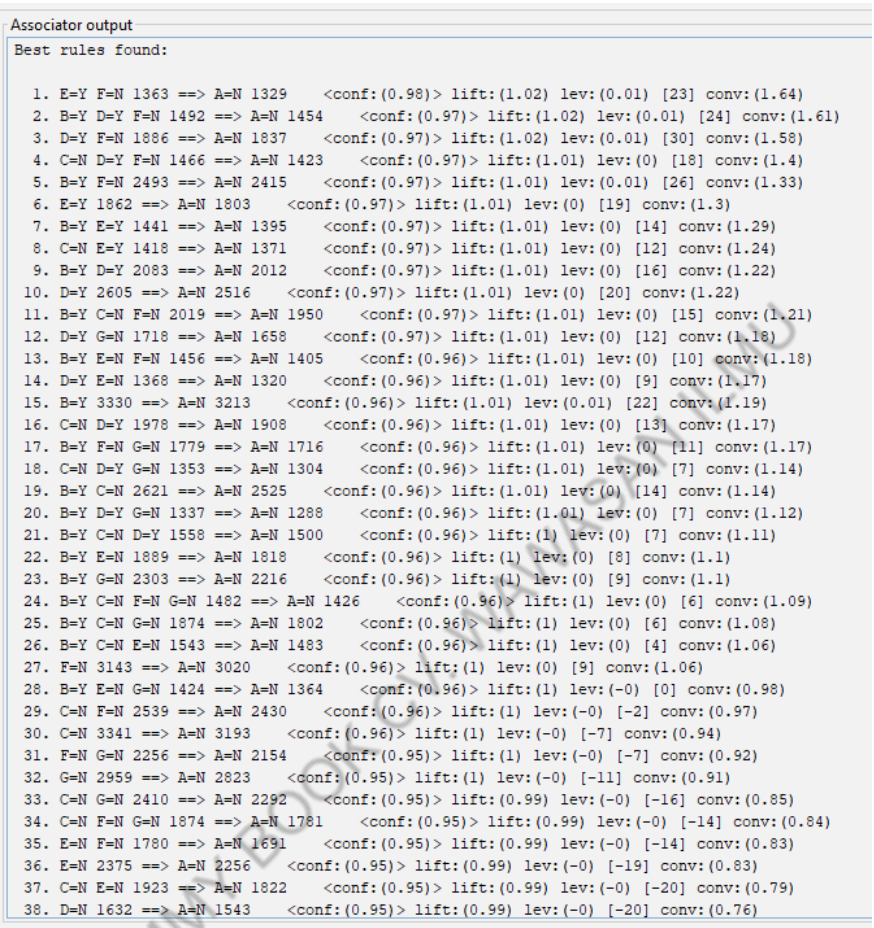
Kode	Atribut
A	grocery misc
B	bread and cake
C	tea
D	biscuits
E	breakfast food
F	coffee
G	jams-spreads

2. Asosiasi Menggunakan Weka

Association Output:

Gambar 5.1 Association Output

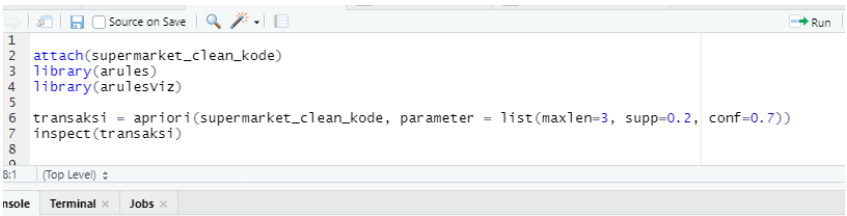
Association Rule Weka:



```
Associator output
Best rules found:

 1. E=Y F=N 1363 ==> A=N 1329    <conf:(0.98)> lift:(1.02) lev:(0.01) [23] conv:(1.64)
 2. B=Y D=Y F=N 1492 ==> A=N 1454    <conf:(0.97)> lift:(1.02) lev:(0.01) [24] conv:(1.61)
 3. D=Y F=N 1886 ==> A=N 1837    <conf:(0.97)> lift:(1.02) lev:(0.01) [30] conv:(1.58)
 4. C=N D=Y F=N 1466 ==> A=N 1423    <conf:(0.97)> lift:(1.01) lev:(0) [18] conv:(1.4)
 5. B=Y F=N 2493 ==> A=N 2415    <conf:(0.97)> lift:(1.01) lev:(0.01) [26] conv:(1.33)
 6. E=Y 1862 ==> A=N 1803    <conf:(0.97)> lift:(1.01) lev:(0) [19] conv:(1.3)
 7. B=Y E=Y 1441 ==> A=N 1395    <conf:(0.97)> lift:(1.01) lev:(0) [14] conv:(1.29)
 8. C=N E=Y 1418 ==> A=N 1371    <conf:(0.97)> lift:(1.01) lev:(0) [12] conv:(1.24)
 9. B=Y D=Y 2083 ==> A=N 2012    <conf:(0.97)> lift:(1.01) lev:(0) [16] conv:(1.22)
10. D=Y 2605 ==> A=N 2516    <conf:(0.97)> lift:(1.01) lev:(0) [20] conv:(1.22)
11. B=Y C=N F=N 2019 ==> A=N 1950    <conf:(0.97)> lift:(1.01) lev:(0) [15] conv:(1.21)
12. D=Y G=N 1718 ==> A=N 1658    <conf:(0.97)> lift:(1.01) lev:(0) [12] conv:(1.18)
13. B=Y E=N F=N 1456 ==> A=N 1405    <conf:(0.96)> lift:(1.01) lev:(0) [10] conv:(1.18)
14. D=Y E=N 1368 ==> A=N 1320    <conf:(0.96)> lift:(1.01) lev:(0) [9] conv:(1.17)
15. B=Y 3330 ==> A=N 3213    <conf:(0.96)> lift:(1.01) lev:(0.01) [22] conv:(1.19)
16. C=N D=Y 1978 ==> A=N 1908    <conf:(0.96)> lift:(1.01) lev:(0) [13] conv:(1.17)
17. B=Y F=N G=N 1779 ==> A=N 1716    <conf:(0.96)> lift:(1.01) lev:(0) [11] conv:(1.17)
18. C=N D=Y G=N 1353 ==> A=N 1304    <conf:(0.96)> lift:(1.01) lev:(0) [7] conv:(1.14)
19. B=Y C=N 2621 ==> A=N 2525    <conf:(0.96)> lift:(1.01) lev:(0) [14] conv:(1.14)
20. B=Y D=Y G=N 1337 ==> A=N 1288    <conf:(0.96)> lift:(1.01) lev:(0) [7] conv:(1.12)
21. B=Y C=N D=Y 1558 ==> A=N 1500    <conf:(0.96)> lift:(1) lev:(0) [7] conv:(1.11)
22. B=Y E=N 1889 ==> A=N 1818    <conf:(0.96)> lift:(1) lev:(0) [8] conv:(1.1)
23. B=Y G=N 2303 ==> A=N 2216    <conf:(0.96)> lift:(1) lev:(0) [9] conv:(1.1)
24. B=Y C=N F=N G=N 1482 ==> A=N 1426    <conf:(0.96)> lift:(1) lev:(0) [6] conv:(1.09)
25. B=Y C=N G=N 1874 ==> A=N 1802    <conf:(0.96)> lift:(1) lev:(0) [6] conv:(1.08)
26. B=Y C=N E=N 1543 ==> A=N 1483    <conf:(0.96)> lift:(1) lev:(0) [4] conv:(1.06)
27. F=N 3143 ==> A=N 3020    <conf:(0.96)> lift:(1) lev:(0) [9] conv:(1.06)
28. B=Y E=N G=N 1424 ==> A=N 1364    <conf:(0.96)> lift:(1) lev:(-0) [0] conv:(0.98)
29. C=N F=N 2539 ==> A=N 2430    <conf:(0.96)> lift:(1) lev:(-0) [-2] conv:(0.97)
30. C=N 3341 ==> A=N 3193    <conf:(0.96)> lift:(1) lev:(-0) [-7] conv:(0.94)
31. F=N G=N 2256 ==> A=N 2154    <conf:(0.95)> lift:(1) lev:(-0) [-7] conv:(0.92)
32. G=N 2959 ==> A=N 2823    <conf:(0.95)> lift:(1) lev:(-0) [-11] conv:(0.91)
33. C=N G=N 2410 ==> A=N 2292    <conf:(0.95)> lift:(0.99) lev:(-0) [-16] conv:(0.85)
34. C=N F=N G=N 1874 ==> A=N 1781    <conf:(0.95)> lift:(0.99) lev:(-0) [-14] conv:(0.84)
35. E=N F=N 1780 ==> A=N 1691    <conf:(0.95)> lift:(0.99) lev:(-0) [-14] conv:(0.83)
36. E=N 2375 ==> A=N 2256    <conf:(0.95)> lift:(0.99) lev:(-0) [-19] conv:(0.83)
37. C=N E=N 1923 ==> A=N 1822    <conf:(0.95)> lift:(0.99) lev:(-0) [-20] conv:(0.79)
38. D=N 1632 ==> A=N 1543    <conf:(0.95)> lift:(0.99) lev:(-0) [-20] conv:(0.76)
```

Gambar 5.2 Association Rule Weka

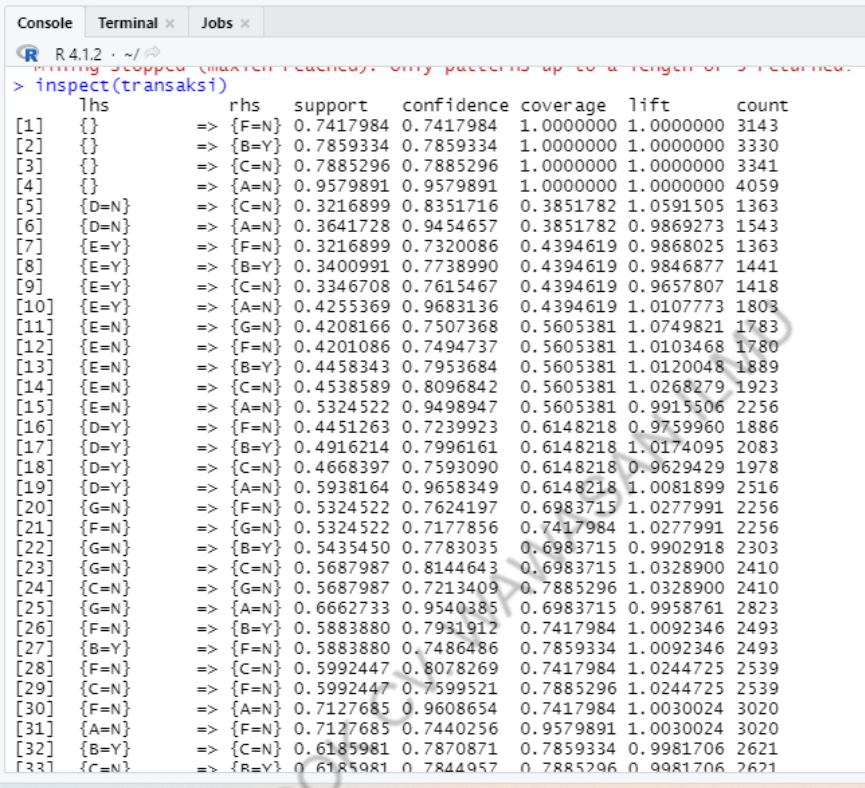
3. Asosiasi Menggunakan R R Code:



```
Source on Save                                                    Run

1
2  attach(supermarket_clean_kode)
3  library(arules)
4  library(arulesviz)
5
6  transaksi = apriori(supermarket_clean_kode, parameter = list(maxlen=3, supp=0.2, conf=0.7))
7  inspect(transaksi)
8
9
8:1  (Top Level)

nsole  Terminal ×  Jobs ×
```

Gambar 5.3 R Code

Association Rule R:



```
Console   Terminal ×   Jobs ×
R 4.1.2 · ~/
> inspect(transaksi)
     lhs          rhs    support   confidence coverage  lift      count
[1]  {}        => {F=N} 0.7417984 0.7417984  1.0000000 1.0000000 3143
[2]  {}        => {B=Y} 0.7859334 0.7859334  1.0000000 1.0000000 3330
[3]  {}        => {C=N} 0.7885296 0.7885296  1.0000000 1.0000000 3341
[4]  {}        => {A=N} 0.9579891 0.9579891  1.0000000 1.0000000 4059
[5]  {D=N}     => {C=N} 0.3216899 0.8351716  0.3851782 1.0591505 1363
[6]  {D=N}     => {A=N} 0.3641728 0.9454657  0.3851782 0.9869273 1543
[7]  {E=Y}     => {F=N} 0.3216899 0.7320086  0.4394619 0.9868025 1363
[8]  {E=Y}     => {B=Y} 0.3400991 0.7738990  0.4394619 0.9846877 1441
[9]  {E=Y}     => {C=N} 0.3346708 0.7615467  0.4394619 0.9657807 1418
[10] {E=Y}     => {A=N} 0.4255369 0.9683136  0.4394619 1.0107773 1803
[11] {E=N}     => {G=N} 0.4208166 0.7507368  0.5605381 1.0749821 1783
[12] {E=N}     => {F=N} 0.4201086 0.7494737  0.5605381 1.0103468 1780
[13] {E=N}     => {B=Y} 0.4458343 0.7953684  0.5605381 1.0120048 1889
[14] {E=N}     => {C=N} 0.4538589 0.8096842  0.5605381 1.0268279 1923
[15] {E=N}     => {A=N} 0.5324522 0.9498947  0.5605381 0.9915506 2256
[16] {D=Y}     => {F=N} 0.4451263 0.7239923  0.6148218 0.9759960 1886
[17] {D=Y}     => {B=Y} 0.4916214 0.7996161  0.6148218 1.0174095 2083
[18] {D=Y}     => {C=N} 0.4668397 0.7593090  0.6148218 0.9629429 1978
[19] {D=Y}     => {A=N} 0.5938164 0.9658349  0.6148218 1.0081899 2516
[20] {G=N}     => {F=N} 0.5324522 0.7624197  0.6983715 1.0277991 2256
[21] {F=N}     => {G=N} 0.5324522 0.7177856  0.7417984 1.0277991 2256
[22] {G=N}     => {B=Y} 0.5435450 0.7783035  0.6983715 0.9902918 2303
[23] {G=N}     => {C=N} 0.5687987 0.8144643  0.6983715 1.0328900 2410
[24] {C=N}     => {G=N} 0.5687987 0.7213409  0.7885296 1.0328900 2410
[25] {G=N}     => {A=N} 0.6662733 0.9540385  0.6983715 0.9958761 2823
[26] {F=N}     => {B=Y} 0.5883880 0.7931912  0.7417984 1.0092346 2493
[27] {B=Y}     => {F=N} 0.5883880 0.7486486  0.7859334 1.0092346 2493
[28] {F=N}     => {C=N} 0.5992447 0.8078269  0.7417984 1.0244725 2539
[29] {C=N}     => {F=N} 0.5992447 0.7599521  0.7885296 1.0244725 2539
[30] {F=N}     => {A=N} 0.7127685 0.9608654  0.7417984 1.0030024 3020
[31] {A=N}     => {F=N} 0.7127685 0.7440256  0.9579891 1.0030024 3020
[32] {B=Y}     => {C=N} 0.6185981 0.7870871  0.7859334 0.9981706 2621
[33] {C=N}     => {B=Y} 0.6185981 0.7844957  0.7885296 0.9981706 2621
```

Gambar 5.4 Association Rule R

4. Hitung Manual

Tabel 5.3 Dataset

Kode	Atribut
A	grocery misc
B	bread and cake
C	tea
D	biscuits
E	breakfast food
F	coffee
G	jams-spreads

Tabel 5.4 ITERASI 2

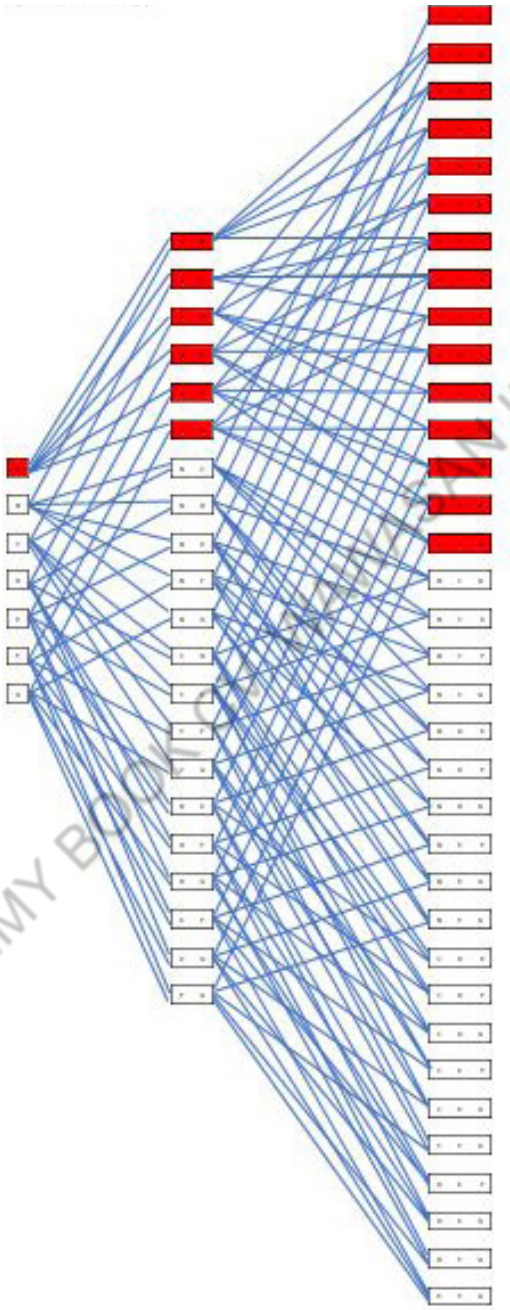
NO	ITEM	TRANSAKSI	SUPPORT (%)	CONFIDENCE
1	A B	117	0.03	0.657
2	A C	30	0.01	0.169
3	A D	89	0.02	0.500
4	A E	59	0.01	0.331
5	A F	55	0.01	0.309
6	A G	42	0.01	0.236
7	B C	709	0.17	0.213
8	B D	2083	0.49	0.626
9	B E	1441	0.34	0.433
10	B F	837	0.20	0.251
11	B G	1027	0.24	0.308
12	C D	627	0.15	0.700
13	C E	444	0.10	0.496
14	C F	292	0.07	0.326
15	C G	347	0.08	0.387
16	D E	1237	0.29	0.475
17	D F	719	0.17	0.276
18	D G	887	0.21	0.340
19	E F	499	0.12	0.268
20	E G	686	0.16	0.368
21	F G	391	0.09	0.357

Tabel 5.5 ITERASI 3

NO	ITEM	TRANSAKSI	SUPPORT (%)	CONFIDENCE
1	A B C	21	0.00	0.18
2	A B D	71	0.02	0.61
3	A B E	46	0.01	0.39
4	A B F	39	0.01	0.33
5	A B G	30	0.01	0.26
6	A C D	19	0.00	0.63

NO	ITEM	TRANSAKSI	SUPPORT (%)	CONFIDENCE
7	A C E	12	0.00	0.40
8	A C F	16	0.00	0.53
9	A C G	12	0.00	0.40
10	A D E	41	0.01	0.46
11	A D F	40	0.01	0.45
12	A D G	29	0.01	0.33
13	A E F	25	0.01	0.42
14	A E G	23	0.01	0.39
15	A F G	21	0.00	0.38
16	B C D	525	0.12	0.74
17	B C E	363	0.09	0.51
18	B C F	235	0.06	0.33
19	B C G	280	0.07	0.39
20	B D E	1020	0.24	0.49
21	B D F	591	0.14	0.28
22	B D G	746	0.18	0.36
23	B E F	404	0.10	0.28
24	B E G	562	0.13	0.39
25	B F G	313	0.07	0.37
26	C D E	330	0.08	0.53
27	C D F	207	0.05	0.33
28	C D G	262	0.06	0.42
29	C E F	142	0.03	0.32
30	C E G	196	0.05	0.44
31	C F G	125	0.03	0.43
32	D E F	371	0.09	0.30
33	D E G	498	0.12	0.40
34	D F G	289	0.07	0.40
35	E FG	221	0.05	0.44

Gambar model



Gambar 5.5 Gambar Model

Tabel 5.6 Atribut

Kode	Atribut	Jumlah	Support
Kode	Atribut	3330	71.97%
A	grocery misc		
B	bread and cake		
C	tea	2795	60.41%
D	biscuits	2605	56.30%
E	breakfast food	788	17.03%
F	coffee	2717	58.72%
G	jams-spreads	4627	

5.4. Studi Kasus

Studi Kasus Asosiasi di Café (Febriani *et. al,*,2021)

Metodologi penelitian menguraikan tahapan penyelesaian masalah berupa langkah-langkah terstruktur pada penelitian. Gambaran alur metodologi penelitian secara rinci dari pengolahan data yang dilakukan ditunjukkan pada gambar berikut.

Gambar 5.6 Flowchart Data Mining Asosiasi

A. Pengumpulan dan Pembersihan Data Mentah Penjualan

Data penjualan yang dikumpulkan diperoleh dari ekstraksi pada sistem kasir Cafe selama 6 bulan berturut-turut sejak September 2020 hingga Maret 2021. File dengan format *Comma Separated Value* (CSV) ini berjumlah sebanyak enam, di mana setiap file berisikan data penjualan masing masing bulan secara berturut-turut di setiap harinya.

1/9/2020 0:00	A200000	2	BTH BEAN LIGHT BIT	FOOD	1	34,000	34,000.00	DEBIT LAIN	LAIN	REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	153	ICE AMER CLASSIC	BEVERAGI	1	20,000	20,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	37	ES KOPI JF SIGNATUI	BEVERAGI	1	18,000	18,000.00	GO FOOD		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	60	ICE TARO MILK BAS	BEVERAGI	1	25,000	25,000.00	GO FOOD		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	65	MILO MACCHIA	BEVERAGI	1	20,000	20,000.00	GO FOOD		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	1	VIVA LA N LIGHT BIT	FOOD	1	27,000	27,000.00	GRAB FOO	FIRMANS'	REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
1/9/2020 0:00	A200000	11	CHEESEBL BURGERS	FOOD	1	32,000	32,000.00	GRAB FOO	FIRMANS'	REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	1	VIVA LA N LIGHT BIT	FOOD	2	27,000	54,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	2	BTH BEAN LIGHT BIT	FOOD	3	34,000	102,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	151	SOUR CRE OPEN FO(	FOOD	3	7,000	21,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	56	ICE RED VI MILK BAS	BEVERAGI	3	25,000	75,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	207	POTATO V SEASONA	FOOD	2	30,000	60,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	128	MIX AND MIX AND	FOOD	3	0	0	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	21	BUTTER FI MIX AND	FOOD	3	9,000	27,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	25	GRILL CHI MIX AND	FOOD	3	22,000	66,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	33	MOZZARE MIX AND	FOOD	3	15,000	45,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	201	GRATIN T OPEN FO(	FOOD	3	4,000	12,000.00	CASH		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	213	GULAI BAI SEASONA	FOOD	3	48,000	144,000.00	TRANSFER		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	201	GRATIN T OPEN FO(	FOOD	3	4,000	12,000.00	TRANSFER		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	14	SPAGHET PASTA AN	FOOD	1	29,000	29,000.00	GO FOOD		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1
2/9/2020 0:00	A200000	11	CHEESEBL BURGERS	FOOD	1	32,000	32,000.00	GO FOOD		REG		0	0	TOTAL :	########	11,091.34	########	Page -1 of 1

Gambar 5.7 Data Mentah Penjualan

Pembersihan data dilakukan pada file setiap bulan, dengan cara yang sama, yaitu penghapusan kolom yang tidak diperlukan seperti kolom yang hanya kalimat yang berulang dan tidak memberi informasi yang relevan. Pembersihan data yang sudah dilakukan kemudian diikuti dengan ditambahnya satu baris di bagian paling atas yang menjadi judul guna memperjelas arti dari data yang ditampilkan pada setiap kolom yang ada. Judul dari masing-masing kolom kemudian diubah ke dalam bahasa yang lebih sederhana agar dapat lebih mudah dipahami. Contohnya adalah perubahan kolom "Date" menjadi "Tanggal", "PLU Name" menjadi "Nama_item" dan "Receipt No" menjadi "Id_Transaksi". Selain itu, juga ditambahkan 2 kolom baru yaitu "Tahun" dan "Bulan" seperti yang ditunjukkan pada gambar 5.8 berikut. Setiap data yang sudah dilakukan pembersihan dan sedikit perubahan, kemudian digabungkan ke dalam 1 file dan berada pada *sheet* yang sama, dengan posisi yaitu diurutkan dari bulan paling awal hingga paling akhir.

Tahun	Bulan	Tanggal	Id_Transaksi	Id_item	Nama_item	Dep_kategori	Group_kategori	Qty_item
2020	9	1/9/2020 0:00	A20000016299	2	8TH BEAN CHCKN WIN	LIGHT BITES	FOOD	1
2020	9	1/9/2020 0:00	A20000016306	153	ICE AMERICANO	CLASSIC	BEVERAGE	1
2020	9	1/9/2020 0:00	A20000016313	37	ES KOPI JADOEL	SIGNATURE	BEVERAGE	1
2020	9	1/9/2020 0:00	A20000016313	60	ICE TARO LATTE	MILK-BASED LATTE	BEVERAGE	1
2020	9	1/9/2020 0:00	A20000016313	65	MILO	MACCHIATO / CHIZU	BEVERAGE	1
2020	9	1/9/2020 0:00	A20000016318	1	VIVA LA NACHOS	LIGHT BITES	FOOD	1
2020	9	1/9/2020 0:00	A20000016318	11	CHEESEBURGER	BURGERS	FOOD	1
2020	9	2/9/2020 0:00	A20000016375	1	VIVA LA NACHOS	LIGHT BITES	FOOD	2
2020	9	2/9/2020 0:00	A20000016375	2	8TH BEAN CHCKN WIN	LIGHT BITES	FOOD	3
2020	9	2/9/2020 0:00	A20000016375	56	ICE RED VELVET LATTE	MILK-BASED LATTE	BEVERAGE	3
2020	9	2/9/2020 0:00	A20000016375	207	POTATO WEDGES	SEASONAL FOOD	FOOD	2
2020	9	2/9/2020 0:00	A20000016375	128	MIX AND MATCH	MIX AND MATCH	FOOD	3
2020	9	2/9/2020 0:00	A20000016375	21	BUTTER FRIED RICE	MIX AND MATCH	FOOD	3
2020	9	2/9/2020 0:00	A20000016375	25	GRILL CHICKEN FILLET	MIX AND MATCH	FOOD	3
2020	9	2/9/2020 0:00	A20000016375	33	MOZZARELLA GRATIN	MIX AND MATCH	FOOD	3
2020	9	2/9/2020 0:00	A20000016376	213	GULAI BAKED RICE	SEASONAL FOOD	FOOD	3
2020	9	2/9/2020 0:00	A20000016388	14	SPAGHETTI BOLOGNE	FRIED RICE AND PASTA	FOOD	1
2020	9	2/9/2020 0:00	A20000016388	11	CHEESEBURGER	BURGERS	FOOD	1
2020	9	2/9/2020 0:00	A20000016388	153	ICE AMERICANO	CLASSIC	BEVERAGE	1
2020	9	2/9/2020 0:00	A20000016388	47	ICE CAFFE LATTE	CLASSIC	BEVERAGE	1
2020	9	2/9/2020 0:00	A20000016392	212	FISH AND CHIPS	SEASONAL FOOD	FOOD	1
2020	9	2/9/2020 0:00	A20000016401	218	SPAGHETTI LUMER	SEASONAL FOOD	FOOD	1
2020	9	2/9/2020 0:00	A20000016401	125	EXTRA MM. MOZZARE	MIX AND MATCH	FOOD	1
2020	9	2/9/2020 0:00	A20000016401	16	SPICY TUNA SPAGHET	FRIED RICE AND PASTA	FOOD	1

Gambar 5.8 Hasil Akhir Pembersihan Data

B. *Association Rule Mining*

Jumlah transaksi penjualan makanan dan minuman di Cafe berturut-turut sejak bulan September 2020 hingga Februari 2021 berdasarkan nama *department* yang menghimpunnya ditampilkan dalam tabel II di bawah. Masing-masing *department* direpresentasikan menggunakan satu buah kode unik berupa huruf kapital dimulai dari huruf A hingga O seperti yang ditunjukkan pada tabel 5.7.

Tabel 5.7 Data Jumlah Transaksi Penjualan

Department	Sep'20	Oct'20	Nov'20	Dec'20	Jan'21	Feb'21
SIGNATURE	40	84	56	30	32	29
CLASSIC	30	66	52	25	35	30
MILK-BASED LATTE	4	9	14	6	3	2
MACCHIATO / CHIZU	2	10	4	4	3	3
BLENDED ICE	0	5	8	5	2	1
CHOCOLATE	11	21	8	11	9	10

Department	Sep'20	Oct'20	Nov'20	Dec'20	Jan'21	Feb'21
MANUAL BREW	0	6	2	2	3	0
MOCKTAILS	17	52	27	16	5	12
SEASONAL DRINK	2	1	9	15	6	6
FRIED RICE AND PASTA	50	71	67	40	23	22
LIGHT BITES	56	77	58	23	28	29
BURGERS	21	18	17	9	13	15
SWEET THINGS	1	7	2	11	10	10
MIX AND MATCH	67	61	49	29	34	23
SEASONAL FOOD		26			11	

Tabel 5.8 Kode yang Digunakan

Department	Kode
SIGNATURE	A
CLASSIC	B
MILK-BASED LATTE	C
MACCHIATO / CHIZU	D
BLENDED ICE	E
CHOCOLATE	F
MANUAL BREW	G
MOCKTAILS	H
SEASONAL DRINK	I
FRIED RICE AND PASTA	J
LIGHT BITES	K
BURGERS	L
SWEET THINGS	M
MIX AND MATCH	N
SEASONAL FOOD	O

Berdasarkan data jumlah transaksi pada tabel 5.8, dapat ditemukan kategori departemen dengan frekuensi transaksi tertinggi pada setiap bulannya. Tabel 5.9 berikut menunjukkan sebanyak 3 buah set item yang merupakan departemen dengan frekuensi tertinggi pada setiap bulannya. Berdasarkan data masing-masing ke-3 buah set item inilah, didapatkan format tabular

numerik seperti ditunjukkan oleh tabel 5.10, yang menjadi dasar untuk pengolahan *data mining* dengan algoritma apriori berikutnya.

Tabel 5.9 Departemen dengan Frekuensi Tertinggi

Bulan	Kode Itemset
Sep'20	N, K, J
Oct'20	A. K. J
Nov'20	J, K, A
Dec'20	J, O, A
Jan'21	B, N, A
Feb'21	B, A, K

Pemberian angka 1 atau 0 sebagai kode numerik memberi makna ada atau tidaknya keberadaan sebuah item pada bulan atau transaksi berikut. Sebagai contoh, pada bulan September 2020 yang diisi dengan itemset N, K, dan J, maka nilai 1 pada bulan tersebut hanya diberi pada kolom N, K, dan J saja. Perlu diketahui bahwa periode bulan pada tabel-tabel berikutnya akan merepresentasikan total keseluruhan transaksi yang ada. Sebagai contoh, bulan "Sep'20" adalah transaksi ke-1, dan "Feb'21" adalah transaksi ke-6. Nilai-nilai pada setiap kolom diakumulasikan ke bawah dan dimasukkan pada baris total. Didapatkan informasi jumlah item N, A, J, B, K, dan O secara berturut-turut sebanyak 2, 5, 4, 2, 4, dan 1 dari total 6 transaksi.

Tabel 5.10 Format Tabular Numerik

Bulan	N	A	J	B	K	O
Sep'20	1	0	1	0	1	0
Oct'20	0	1	1	0	1	0
Nov'20	0	1	1	0	1	0
Dec'20	0	1	1	0	0	1
Jan'21	1	1	0	1	0	0
Feb'21	0	1	0	1	1	0
Total	2	5	4	2	4	1

Tahap selanjutnya yaitu mencari kombinasi item yang memenuhi syarat minimum dari nilai *support* yang telah ditentukan, yaitu sebesar 0.3. Perhitungan dilakukan dimulai dari 1-itemset berlanjut hingga k-itemset, yaitu saat iterasi selesai sehingga pembentukan itemset tidak dapat dilakukan lagi.

1) Pembentukan *frequent* 1-itemset

Dasar perhitungan pada pembentukan 1-itemset menggunakan total dari masing-masing item keseluruhan transaksi tabel 5.10 sebelumnya. Perhitungan nilai *support* 1-itemset adalah sebagai berikut.

$$\text{Support (N)} = \frac{\sum \text{Jumlah transaksi yang mengandung N}}{\text{Total transaksi}}$$

$$\text{Support (N)} = \frac{2}{6} = 0{,}333 \qquad \text{Support (A)} = \frac{5}{6} = 0{,}833$$

$$\text{Support (J)} = \frac{4}{6} = 0{,}667 \qquad \text{Support (B)} = \frac{2}{6} = 0{,}333$$

$$\text{Support (K)} = \frac{4}{6} = 0{,}667 \qquad \text{Support (O)} = \frac{1}{6} = 0{,}167$$

Rangkuman hasil dari seluruh perhitungan nilai *support* final pada 1-itemset ditunjukkan oleh tabel 5.11. Berdasarkan perhitungan di atas, diketahui bahwa itemset O memiliki nilai *support* 0.167, yang berada di bawah minimum *support*. Oleh karena itu, itemset O harus dihilangkan dan tidak akan dilanjutkan ke perhitungan selanjutnya.

Tabel 5.11 Support 1-Itemset Final

Temset	Support
N	0.333
A	0.833
J	0.667
B	0.333
K	0.667

2) Pembentukan *frequent* 2-itemset

Pembentukan 2-itemset menggunakan kombinasi dari itemset final yang didapat pada pola pembentukan 1-itemset. Iterasi dari pembentukan kombinasi itemset ini yaitu menentukan keberadaan itemset, yang kemudian diberi tanda "Y/N" dimana Y mewakili *yes*, dan N mewakili *no*. Setelah itu, banyak nilai "Y" pada kolom tersebut dijumlahkan sebagai total dari kombinasi 2-itemset yang terbentuk. Iterasi pembentukan 2-itemset berjumlah 10, dikarenakan

pola calon yang terbentuk berjumlah 10. Adapun calon himpunan kombinasi 2-itemset yang dimaksud yaitu {N,A}, {N,J}, {N,B}, {N,K}, {A,J}, {A,B}, {A,K}, {J,B}, {J,K}, {B,K}.

Total dari masing-masing kombinasi berdasarkan proses iterasi yang sudah dilakukan ditunjukkan pada tabel 5.12 berikut.

Tabel 5.12 Total Calon 2-Itemset

Itemset	Total
N, A	1
N, J	1
N, B	1
N, K	1
A, J	3
A, B	2
A, K	3
J, B	0
J, K	3
B, K	1

Berdasarkan perhitungan nilai penunjang (*support*), kombinasi 2-itemset final yang terbentuk ditunjukkan secara ringkas pada tabel berikut.

Tabel 5.13 Support 2-Itemset Final

Itemset	**Support**
A, J	0.500
A, B	0.333
A, K	0.500
J, K	0.500

Adapun rincian perhitungan *support* dari masing-masing kombinasi itemset adalah sebagai berikut:

$$\text{Support (A, J)} = \frac{\sum \text{Jumlah transaksi mengandung A dan J}}{\text{Total transaksi}}$$

$$\text{Support (A, J)} = \frac{3}{6} = 0{,}500$$

$$\text{Support (A, B)} = \frac{2}{6} = 0{,}333$$

$$\text{Support (A, K)} = \frac{3}{6} = 0{,}500$$

$$\text{Support (J, K)} = \frac{3}{6} = 0{,}500$$

3) Pembentukan *frequent* 3-itemset

Kombinasi itemset dalam F2 kemudian digabungkan membentuk calon 3-itemset. Penggabungan itemset dilihat dari adanya kesamaan pada k-itemset pertama. Itemset-itemset yang dapat digabungkan adalah itemset-itemset yang memiliki kesamaan dalam k-1 item pertama. Itemset himpunan {A,J} dan {A,B} memiliki kesamaan pada {A}, sehingga membentuk calon 3-itemset {A,J,B}. Itemset himpunan {A,J} dan {A,K} memiliki kesamaan pada {A}, sehingga membentuk calon 3-itemset {A,J,K}. Itemset himpunan {A,B} dan {A,K} memiliki kesamaan pada {A}, sehingga membentuk calon 3-itemset {A,B,K}.

Iterasi pembentukan 3-itemset pada {A,J,B}, {A,J,K} dan {A,B,K} secara berturut-turut memiliki total 0, 2, dan 1. Minimum *support* 0.3 hanya dipenuhi oleh kombinasi {A,B,K} dengan *support* sebesar 0.333, dengan perhitungan sebagai berikut.

$$\text{Support (A, J, K)} = \frac{\sum \text{Jumlah transaksi mengandung A, J, K}}{\text{Total transaksi}}$$

$$\text{Support (A, J, K)} = \frac{2}{6} = 0{,}333$$

Tabel 5.14 Support 3-Itemset Final

Itemset	Support
A, J, K	0.333

Pembentukan kombinasi *frequent* itemset berakhir pada kombinasi 3-itemset. Hal ini dikarenakan itemset yang terbentuk pada 3-itemset berjumlah 1 himpunan yaitu {A,J,K} dan tidak dapat dibentuk calon itemset baru. Maka dari itu, pembentukan aturan

asosiasi dapat dilakukan dengan menggabungkan itemset-itemset yang sudah terbentuk dan memenuhi aturan *support* 0.3 menjadi sebuah aturan *if x, then y* yang berarti, apabila x dibeli, maka y juga akan dibeli. Pembentukan aturan-aturan asosiasi awal dapat dilihat pada tabel 5.15 di bawah. Ke-18 aturan tersebut, dihitung nilai *confidence*-nya. *Confidence* atau nilai kepastian menunjukkan persentase kuatnya hubungan antar item dalam aturan asosiasi. Perhitungan *confidence* dari seluruh aturan asosiasi adalah sebagai berikut:

$$\text{Confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{Support (A U B)}}{\text{Support (A)}}$$

$$\text{Conf } (A \Rightarrow J) = \frac{0{,}500}{0,\ 833} = 0{,}600 \qquad \text{Conf } (J \Rightarrow A) = \frac{0{,}500}{0{,}667} = 0{,}750$$

$$\text{Conf } (A \Rightarrow B) = \frac{0{,}333}{0,\ 833} = 0{,}400 \qquad \text{Conf } (B \Rightarrow A) = \frac{0{,}333}{0,\ 333} = 1{,}000$$

$$\text{Conf } (A \Rightarrow K) = \frac{0{,}500}{0,\ 833} = 0{,}600 \qquad \text{Conf } (K \Rightarrow A) = \frac{0{,}500}{0{,}667} = 0{,}600$$

$$\text{Conf } (J \Rightarrow K) = \frac{0{,}500}{0,\ 667} = 0{,}750 \qquad \text{Conf } (K \Rightarrow J) = \frac{0{,}500}{0{,}667} = 0{,}600$$

$$\text{Conf } (A, J \Rightarrow K) = \frac{0{,}333}{0{,}500} = 0{,}667$$

$$\text{Conf } (A, K \Rightarrow J) = \frac{0{,}333}{0{,}500} = 0{,}667$$

$$\text{Conf } (J, K \Rightarrow A) = \frac{0{,}333}{0{,}500} = 0{,}667$$

Tabel 5.15 Pembentukan Aturan Asosiasi

Aturan	Support	Confidence
Jika membeli A, maka akan membeli J	0.500	0.600
Jika membeli J, maka akan membeli A	0.500	0.750
Jika membeli A, maka akan membeli B	0.333	0.400

Aturan	Support	Confidence
Jika membeli B, maka akan membeli A	0.333	1.000
Jika membeli A, maka akan membeli K	0.500	0.600
Jika membeli K, maka akan membeli A	0.500	0.750
Jika membeli J, maka akan membeli K	0.500	0.750
Jika membeli K, maka akan membeli J	0.500	0.750
Jika membeli A dan J, maka akan membeli K	0.333	0.667
Jika membeli A dan K, maka akan membeli J	0.333	0.667
Jika membeli J dan K, maka akan membeli A	0.333	0.667

Hasil dari perhitungan nilai *confidence* menyisakan 5 dari 18 aturan asosiasi yang terbentuk. Dengan minimum *confidence* sebesar 0.7, aturan-aturan asosiasi yang memenuhi standar ditunjukkan pada tabel XI di bawah. Untuk memastikan validitas dari aturan yang terbentuk, dilakukan perhitungan *lift ratio*. Perhitungan nilai *lift* ini didapat dengan membagi nilai *confidence* aturan yang terbentuk dengan nilai *confidence benchmark*-nya.

Tabel 5.16 Perhitungan Nilai Lift

Aturan	Support	Conf.	Lift
Jika membeli J, maka akan membeli A	0.500	0.750	0.900
Jika membeli B, maka akan membeli A	0.333	1.000	1.200
Jika membeli K, maka akan membeli A	0.500	0.750	0.900
Jika membeli J, maka akan membeli K	0.500	0.750	1.125
Jika membeli K, maka akan membeli J	0.500	0.750	1.125

Perhitungan nilai *lift* dari masing-masing aturan asosiasi:

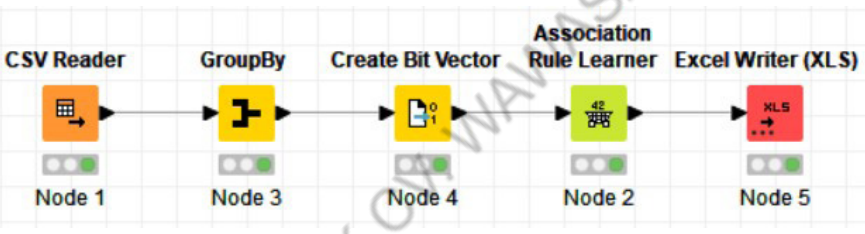
$$\text{Lift } (J \Rightarrow A) = \frac{0{,}750}{0{,}833} = 0{,}900 \qquad \text{Lift } (B \Rightarrow A) = \frac{1{,}000}{0{,}833} = 1{,}200$$

$$\text{Lift } (K \Rightarrow A) = \frac{0{,}750}{0{,}833} = 0{,}900 \qquad \text{Lift } (J \Rightarrow K) = \frac{0{,}750}{0{,}667} = 1{,}125$$

$$\text{Lift } (K \Rightarrow J) = \frac{0{,}750}{0{,}667} = 1{,}125$$

Nilai *lift* yang didapat dari perhitungan di atas, terbagi ke dalam dua jenis, yaitu *lift* dengan nilai kurang dari 1, dan *lift* dengan nilai lebih dari 1. *Lift* bernilai kurang dari 1, menandakan bahwa aturan asosiasi terkait tidak valid dikarenakan kemunculan item x berkolerasi negatif dengan kemunculan item y. Oleh sebab itu, aturan tersebut sebaiknya dihilangkan. Dengan menyingkirkan aturan asosiasi {J ⇒ A} dan {K ⇒ A}, didapatkan aturan asosiasi final diantaranya adalah {B ⇒ A}, {J ⇒ K}, {K ⇒ J}.

Selain melakukan perhitungan secara manual menggunakan rumus yang ada, proses *data mining* kembali dilakukan menggunakan bantuan *software* untuk mengetahui *association rule* yang terbentuk. *Software* yang digunakan yaitu Konstanz Information Miner (KNIME). Data awal yang digunakan sebagai input yaitu data pada tabel V di atas yang diubah ke dalam format kategorik Y dan N, dimana angka 1 berarti Y, dan angka 0 berarti N.



Gambar 5.9 Rancangan Pembentukan Aturan Asosiasi KNIME

Hasil akhir dari penggunaan KNIME yaitu ditemukannya sebanyak 23 aturan asosiasi. Ke- 23 aturan asosiasi yang terbentuk beserta nilai *support*, *confidence*, dan *lift*-nya dapat dilihat pada gambar 17 di bawah. Penelitian ini menggunakan minimum *support* sebesar 0.3, dan minimum *confidence* sebesar 0.7. Maka, pembentukan aturan asosiasi yang tidak memenuhi standar dibuang. Berdasarkan ke-23 *rule* yang terbentuk, *rule* yang memenuhi minimum *support* dan *confidence* yaitu "rule 17", dengan nilai *support* sebesar 0.4, *confidence* 1, dan *lift* 1.25. *Rule* ini menyatakan "Jika membeli B, maka akan membeli A" atau {B ⇒ A}. Aturan tersebut sama dengan salah satu aturan yang terbentuk pada perhitungan manual yang dapat dilihat kembali pada tabel 5.16 di atas.

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.2	0.5	1.25	N	<---	[A,B]
rule1	0.2	1	1.25	A	<---	[B,N]
rule2	0.2	1	2.5	B	<---	[A,N]
rule3	0.2	0.5	1.25	N	<---	[J,K]
rule4	0.2	1	1.667	J	<---	[K,N]
rule5	0.2	1	1.667	K	<---	[J,N]
rule6	0.2	0.5	0.625	A	<---	[J,K]
rule7	0.2	0.5	0.833	J	<---	[A,K]
rule8	0.2	0.5	0.833	K	<---	[A,J]
rule9	0.2	1	1.25	A	<---	[J,O]
rule10	0.2	1	1.667	J	<---	[A,O]
rule11	0.2	0.5	2.5	O	<---	[A,J]
rule12	0.2	1	1.25	A	<---	[B,K]
rule13	0.2	0.5	1.25	B	<---	[A,K]
rule14	0.2	0.5	0.833	K	<---	[A,B]
rule15	0.4	0.667	0.833	A	<---	[J]
rule16	0.4	0.5	0.833	J	<---	[A]
rule17	0.4	1	1.25	A	<---	[B]
rule18	0.4	0.5	1.25	B	<---	[A]
rule19	0.4	0.667	0.833	A	<---	[K]
rule20	0.4	0.5	0.833	K	<---	[A]
rule21	0.4	0.667	1.111	J	<---	[K]
rule22	0.4	0.667	1.111	K	<---	[J]

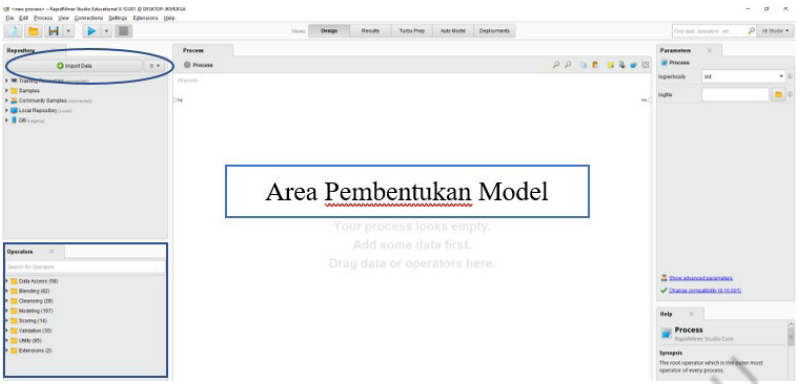
Gambar 5.10 Output Asosiasi KNIME

Ditemukan 3 aturan asosiasi yaitu {CLASSIC ⇒ SIGNATURE}, {FRIED RICE AND PASTA ⇒ LIGHT BITES}, {LIGHTBITES ⇒ FRIED RICE AND PASTA} dari data mining transaksi penjualan. Kombinasi dari menu classic dan signature, serta fried rice and pasta dan light bites dapat dijadikan sebagai salah satu strategi penjualan. Usulan utama yang diberikan untuk perusahaan yaitu agar melakukan pembenahan terkait penyimpanan dan pengelolaan data dengan melakukan transformasi ke dalam ruang lingkup digital.

Studi Kasus pada Customer Display Product (Alfiandi,2021)

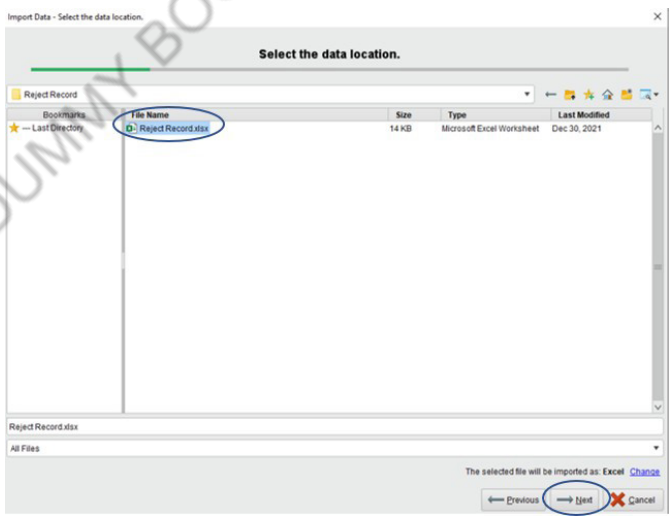
Membangun sebuah model dari *set data* merupakan proses yang dilakukan berulang hingga akhirnya didapat sebuah model yang sesuai dan parameter yang optimal. Pengerjaan model *building* pada penelitian ini menggunakan *software* RapidMiner, hal ini dikarenakan pada *software* ini memiliki berbagai *operator* yang dapat digunakan menurut algoritma *data mining* yang akan digunakan. Berikut ini akan dijelaskan langkah untuk membangun model asosiasi untuk produk CDP dengan menggunakan *Software* RapidMiner.

Setelah *Software* RapidMiner sudah dibuka, akan muncul tampilan *welcome to RapidMiner Studio*. Pada tampilan ini pilih *start with blank process*. Fungsi ini digunakan untuk membuat proses baru dengan tampilan pada area pembentukan model.



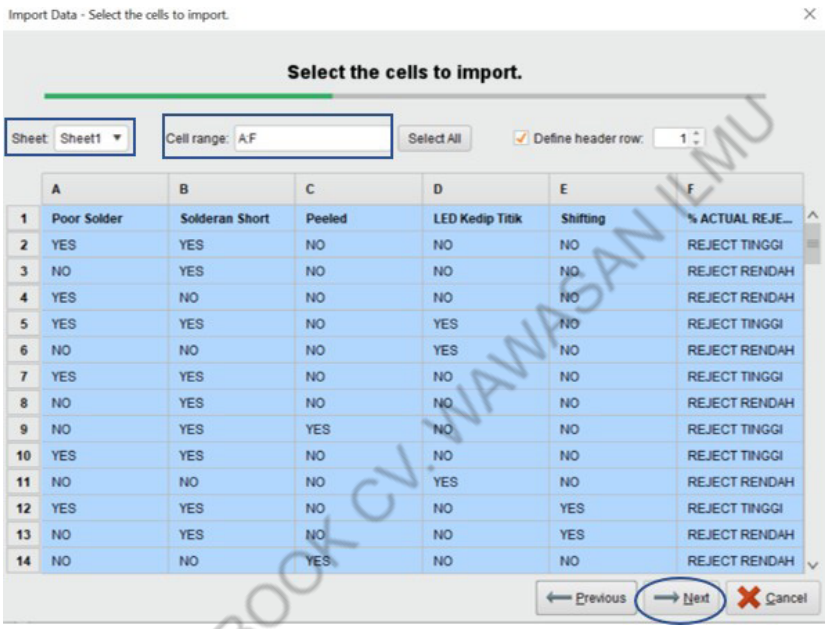
Gambar 5.11 Tampilan Awal Pembentukan Model *Software* RapidMiner

Selanjutnya akan muncul tampilan awal Pembentukan Model *Software* RapidMiner. Pada tampilan ini pada area tengah terdapat area Process. Area *process* ini dapat membentuk model yang terdiri atas *input* berupa setdata, operator fungsional, dan juga terhubung ke *output* pemodelan. Pada operator fungsional terdapat pada bagian kanan bawah, dan dapat dipilih node fungsional sesuai dengan teknik *Data mining* yang digunakan. Pada pembuatan model ini langkah pertama yang dilakukan yaitu *import* data, dengan cara memilih "+ *Import Data*" yang berada pada *tab Repository* yang berada pada kiri atas.



Gambar 5.12 Pemilihan Set data

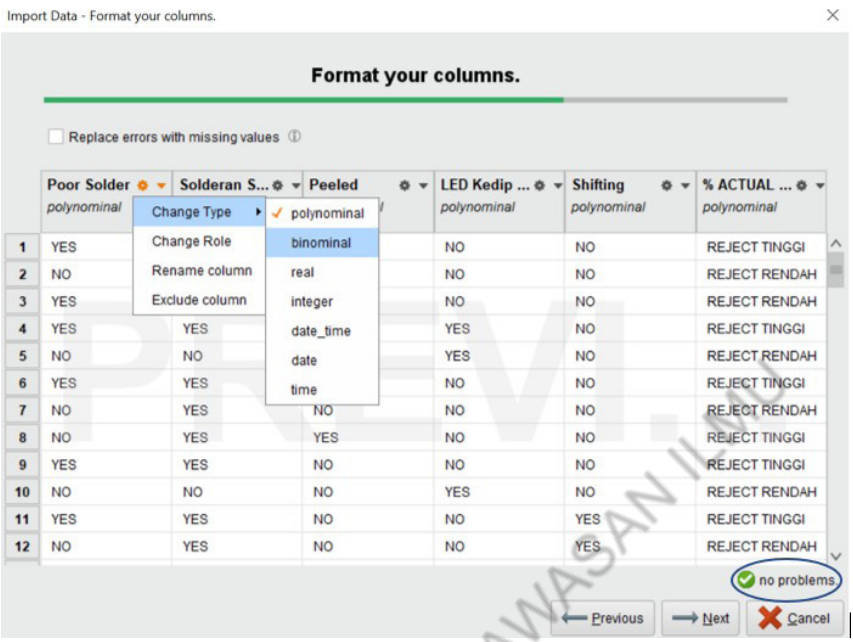
Berikutnya *step* yang dilakukan adalah pemilihan data berdasarkan lokasinya. Pada pemilihan data ini, jenis *files* yang dapat digunakan adalah *file* berjenis xlsx, ataupun csv. Jenis *file* tersebut dapat dihasilkan oleh *software Microsoft Excel*. Pada Gambar pemilihan setdata, terlihat bahwa setdata yang digunakan adalah data *Reject Record* berjenis file xlsx. Langkah selanjutnya klik tombol "*Next*" yang berada pada kanan bawah.



	A	B	C	D	E	F
1	Poor Solder	Solderan Short	Peeled	LED Kedip Titik	Shifting	% ACTUAL REJE...
2	YES	YES	NO	NO	NO	REJECT TINGGI
3	NO	YES	NO	NO	NO	REJECT RENDAH
4	YES	NO	NO	NO	NO	REJECT RENDAH
5	YES	YES	NO	YES	NO	REJECT TINGGI
6	NO	NO	NO	YES	NO	REJECT RENDAH
7	YES	YES	NO	NO	NO	REJECT TINGGI
8	NO	YES	NO	NO	NO	REJECT RENDAH
9	NO	YES	YES	NO	NO	REJECT TINGGI
10	YES	YES	NO	NO	NO	REJECT TINGGI
11	NO	NO	NO	YES	NO	REJECT RENDAH
12	YES	YES	NO	NO	YES	REJECT TINGGI
13	NO	YES	NO	NO	YES	REJECT RENDAH
14	NO	NO	YES	NO	NO	REJECT RENDAH

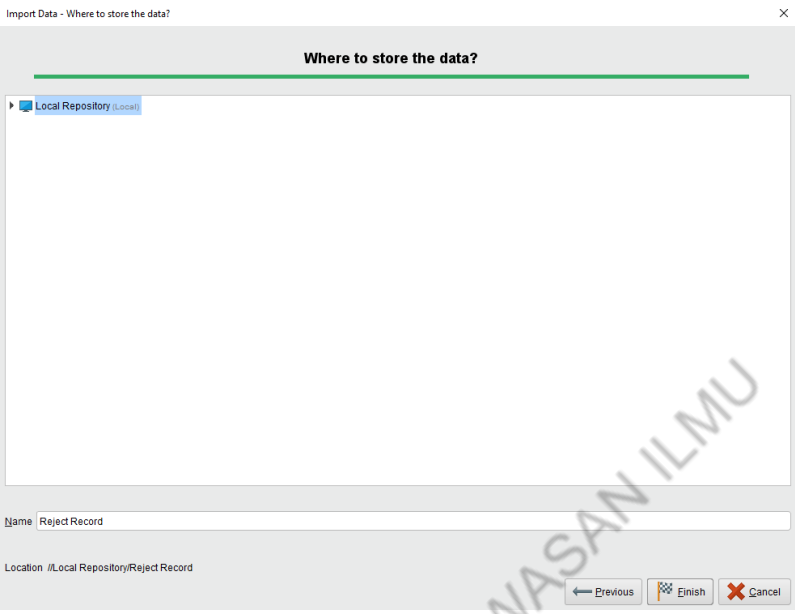
Gambar 5.13 Pemilihan Data yang Akan Digunakan

Pada Gambar terdapat tampilan untuk pemilihan data yang akan digunakan. Pada setdata yang digunakan hanya terdapat 1 halaman sehingga pada pemilihan halaman pada Sheet1, pemilihan halaman ini terdapat pada kolom kanan atas. Untuk pemilihan kolom yang digunakan merupakan kolom A hingga kolom F sehingga pada "CELL RANGE" akan terisi A:F. Selanjutnya apabila pemilihan data selesai dilakukan dapat dilanjutkan dengan klik tombol "*Next*" yang berada pada kanan bawah.



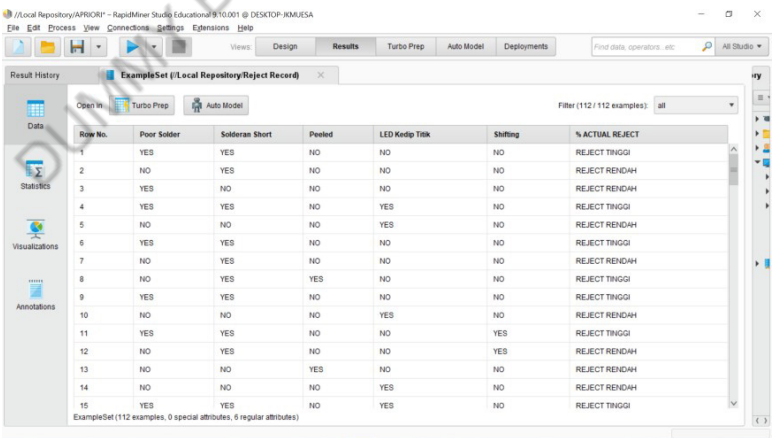
Gambar 5.14 Pengaturan Tipe Data

Langkah selanjutnya yaitu mengatur tipe data. Pada set data *Reject Record* memiliki jenis atribut berjumlah 20. Dimana 5 jenis kecacatan pada modul yang terdapat pada produk CDP, seperti solderan *Short*, LED kedip titik, *Poor solder*, *peeled*, *shifting* dan 1 atribut yang merupakan *label* pada kolom terakhir yaitu % *Actual Reject* sebagai penentu *reject* tinggi atau rendah. Pada setiap kolom, sebelumnya memiliki tipe data *polynominal*, sehingga harus diubah kedalam tipe data *binominal*. Hal ini dikarenakan pada setiap kolom hanya mempunyai dua jenis respon yaitu "YES" atau "NO". Selanjutnya apabila pada data tidak ditemukan masalah akan muncul *text* "no problems", dan dapat dilanjutkan dengan mengklik perintah "Next".



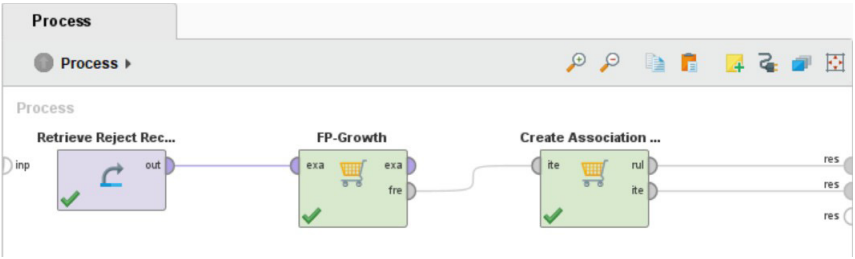
Gambar 5.15 Pengaturan Penyimpanan Data

Langkah terakhir adalah pengaturan untuk penyimpanan data. Penyimpanan ini dilakukan sehingga setdata yang sudah diimport dapat digunakan kembali sehingga proses import pada data yang sama tidak dilakukan berulangkali. Selanjutnya klik "finish", dan apabila data berhasil diimport maka tampilan *software* RapidMiner akan terlihat seperti pada gambar 5.16.



Gambar 5.16 Tampilan Data Selesai Dimasukan ke Software RapidMiner

- Pembuatan Algoritma APRIORI



Gambar 5.17 Rangkaian Proses Pemodelan Data

Pada Gambar menunjukan alur proses pemodelan data yang dibangun dengan memasukkan operator satu per satu. Setiap operator memiliki node yang dapat disambungkan dengan node dari operator lainnya agar fungsi dapat berjalan dan menghasilkan keluaran yang diinginkan. Dapat terlihat model bersumber dari *input* data *Reject Record* pada bagian kiri dan akan menghasilkan *output data mining* pada bagian kanan. Berikut ini merupakan tahapan yang dilakukan dan penjelasan untuk setiap operator yang digunakan:

1. *Retrieve Reject Record*

Pada operator *Retrieve Reject Record* adalah operator yang digunakan untuk membaca data *reject record* yang terdapat pada *data repository*. Fungsi operator ini merupakan *input* proses pada pemodelan. Node *out* dihubungkan dengan operator berikutnya.

2. *FP Growth*

FP-Growth atau *Frequent Pattern – Growth* merupakan operator yang berfungsi untuk dapat menjalankan teknik aturan asosiasi. Operator ini akan menghitung kombinasi dari adanya kemunculan *itemsets* dari setdata. Untuk menjalankan operator ini terdapat aturan yaitu data bertipe binominal atau hanya terdiri dari dua keputusan, seperti "YA" dan "TIDAK", atau "*Reject* rendah" dan "*Reject* tinggi".

3. *Create Association*

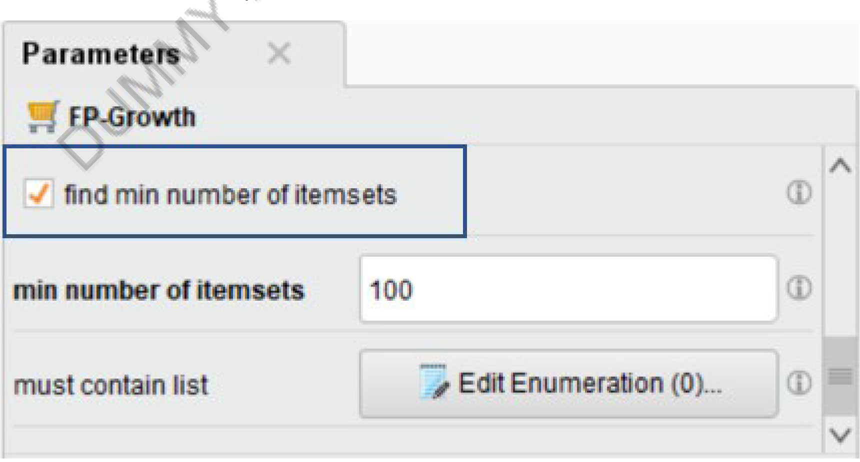
Hasil dari operator sebelumnya yaitu FP-Growth menghasilkan kombinasi *frequent itemsets* dari seluruh data, sehingga memiliki jumlah yang cukup banyak sehingga tidak dapat membuat kesimpulan. Sehingga *Create Association Rules* perlu digunakan

sebagai operator untuk menghasilkan *output* aturan asosiasi atau pola hubungan dari *frequent itemsets* yang telah terbentuk.

- Output Teknik Asosiasi

Software RapidMiner akan mengolah data sehingga akan didapat *Frequent Itemset. Frequent Itemset* merupakan sekumpulan itemset yang memiliki kemunculan bersama. Dimana pada kombinasi Itemset yang ada memiliki nilai *support* yang sama atau lebih besar dari nilai minimum batas *support*. Itemset merupakan kumpulan dari satu atau lebih item, dimana item merupakan sebuah atribut. Dimana pada penilitian ini, itemset merupakan hubungan antara 5 item jenis kecacatan yaitu *Poor* Solder, solderan shot, peeled, LED kedip titik dan shifting. Dan nilai support merupakan frekuensi dari kemunculan itemset. Sehingga fungsi dari penggunaan algoritma asosiasi adalah untuk menemukan kombinasi dari beberapa item yang muncul bersama pada nilai *support* tertentu.

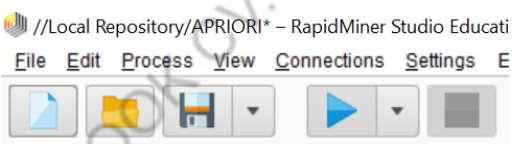
Pada penggunaan dari algoritma ini dibutuhkan nilai minimum support, untuk mengetahui apakah nilai support dari itemset dapat terpenuhi, dan apabila terpenuhi maka itemset tersebut akan terpilih oleh RapidMiner. Untuk mencari nilai minimum batas support dapat dicari dengan perhitungan antara banyaknya itemset yang muncul dibagi total banyaknya data, akan tetapi pada *software* RapidMiner memiliki pilihan "*Find Min Number of Itemset*", sehingga dengan pilihan opsi tersebut maka nilai minimum *support* tidak dibutuhkan. Berikut ini Gambar 5.18 mengenai opsi pilihan "*Find Min Number of Itemset*".



Gambar 5.18 Pilihan opsi *find min number of itemsets*

Terkait nilai minimum *support* tidak dibutuhkan dikarenakan pada RapidMiner akan mencari kumpulan item dan membuat itemsets dari setdata dengan menggunakan jumlah nilai *support* terbesar, selanjutnya dari itemsets tersebut otomatis terpilih menjadi frequent itemsets. Sehingga didapatkan nilai minimum *confidence* yaitu sebesar 0,567. Sehingga itemset yang muncul hanya itemset yang memenuhi syarat, yaitu memiliki nilai *confidence* lebih dari nilai *minimum* confidence 0,567.

Pada proses model yang terdapat pada gambar 5.19 memiliki operator "FP- *Growth*" dan "*Create Association Rules*". Hal ini dikarenakan fungsi adanya operator FP-*Growth* untuk mendapatkan nilai *frequent itemsets,* dan tidak memiliki aturan asosiasi. Sehingga dengan adanya operator *Create Association rules* dapat menghasilkan aturan asosiasi berbentuk *If-Then Rules* atau aturan jika-maka, untuk memperlihatkan hubungan antara jenis kecacatan dan juga nilai *confidence* dari hubungan tersebut, dengan logika apabila 1 jenis kecacatan muncul, apakah ada kecacatan yang muncul secara bersamaan dan dapat menjadi faktor adanya *reject* yang tinggi. Selanjutnya apabila model sudah dibuat, dapat dijalankan atau *Run* seperti pada Gambar 5.19 berikut.



Gambar 5.19 Opsi Run pada RapidMiner

Setelah proses dijalankan maka akan menghasilkan *output* berupa *Association rules* seperti berikut:

```
Association Rules
1. [% ACTUAL REJECT] --> [Solderan Short, Poor
   Solder] (confidence: 0.567)
2. [Solderan Short] --> [% ACTUAL REJECT, Poor
   Solder] (confidence: 0.594)
3. [Solderan Short] --> [Poor Solder] (confidence:
   0.656)
4. [Poor Solder] --> [% ACTUAL REJECT, Solderan
   Short] (confidence: 0.667)
5. [% ACTUAL REJECT] --> [Poor Solder] (confidence:
   0.716)
```

6. [% ACTUAL *REJECT*, Solderan *Short*] --> [*Poor* Solder] (confidence: 0.731)
7. [*Poor* Solder] --> [Solderan *Short*] (confidence: 0.737)
8. [% ACTUAL *REJECT*] --> [Solderan *Short*] (confidence: 0.776)
9. [% ACTUAL *REJECT*, *Poor* Solder] --> [Solderan *Short*] (confidence: 0.792)
10. [Solderan *Short*] --> [% ACTUAL *REJECT*] (confidence: 0.812)
11. [*Poor* Solder] --> [% ACTUAL *REJECT*] (confidence: 0.842)
12. [Solderan *Short*, *Poor* Solder] --> [% ACTUAL *REJECT*] (confidence: 0.905)

Nilai *support* merupakan nilai untuk menentukan kemungkinan suatu *item* muncul bersama dengan item lain, dan pada parameter *confidence* menentukan seberapa sering aturan *if-Then Rules* muncul. Berdasarkan hasil aturan asosiasi dihasilkan 12 aturan asosiasi dari 56 item atribut yaitu 5 jenis cacat, dan 1 atribut keputusan *reject.* Selain itu, hasil aturan tersebut memiliki pola kombinasi 2 atribut kecacatan dominan, yaitu *Poor* Solder dan Solderan *Short* dan berasosiasi dengan item % *Actual Reject.* Pada nilai *confidence* yang didapatkan memiliki nilai terkecil sebesar 0,567, dan nilai *confidence* terbesar yaitu 0,905. Sehingga didapatkan kesimpulan apabila muncul jenis kecacatan *Poor solder* maka akan muncul juga jenis kecacatan *solderan Short*, dan akan mengakibatkan *reject* yang tinggi.

Studi Kasus Asosiasi Data Mining pada perusahaan penerbangan

Data–data yang digunakan pada pengolahan data *mining* menggunakan data hasil kuesioner tingkat kepuasan pada penerbangan Jakarta – Surabaya dan Surabaya – Jakarta yang telah mengalami pengolahan berupa perubahan kode tingkat kepuasan menjadi tingkat kepuasan.

1. Penerbangan Jakarta – Surabaya

Tahap awal dalam perhitungan manual data mining menggunakan metode bayes adalah dengan cara menentukan probabilitas posterior klasifikasi terlebih dahulu. Probabilitas prior didapat dengan cara pembagian jumlah salah satu klasifikasi dengan jumlah koresponden, atau dapat dilihat pada rumus 5.5.

Probabilitas Prior = $\frac{x_i}{\sum_{i=1}^{n} x}$ .....(5.5)

Pada penerbangan Jakarta – Surabaya memiliki probabilitas posterior yang memiliki nilai besar pada klasifikasi puas dengan nilai diatas 0,5. Data hasil probabilitas posterior selengkapnya dapat dilihat pada tabel 5.17.

Tabel 5.17 Probabilitas Prior Penerbangan Jakarta – Surabaya

Klasifikasi	Count	Probabilitas Prior
Sama Puas	52	0,390977444
Puas	74	0,556390977
Netral	6	0,045112782
Tidak Puas	0	0
Sama Sekali Tidak Puas	1	0,007518797
Total	133	1

Tahap selanjutnya setelah mendapatkan probabilitas prior adalah mendapatkan probabilitas interaksi antara tiap atribut dengan klasifikasi. Probabilitas interaksi atribut dengan klasifikasi dapat dilihat pada tabel berikutnya.

Terlihat pada tabel 5.18 probabilitas dari interaksi antara atribut *Website Service* dengan klasifikasi. Pada saat *Website Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,75 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.18 Probabilitas Interaksi *Website Service* Penerbangan Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Website Service	Sama Puas	0,754098361	0,229508197	0,016393443			1
	Puas	0,105263158	0,859649123	0,035087719			1
	Netral		0,769230769	0,230769231			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.19 probabilitas dari interaksi antara atribut *Contact Center Service* dengan klasifikasi. Pada saat *Contact Center Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,8 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.19 Probabilitas Interaksi *Contact Center Service* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Contact Center Service	Sama Puas	0,803571429	0,196428571				1
	Puas	0,109375	0,875	0,015625			1
	Netral		0,583333333	0,416666667			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.20 probabilitas dari kemungkinan antara atribut *Sales Office Service* dengan klasifikasi. Pada saat *Sales Office Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,75 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.20 Probabilitas Interaksi *Sales Office Service* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Sales Office Service	Sama Puas	0,754385965	0,245614035				1
	Puas	0,15	0,8	0,05			1
	Netral		0,8	0,2			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.21 probabilitas dari interaksi antara atribut *Check-in Service* dengan klasifikasi. Pada saat *Check-in Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,66 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.21 Probabilitas Interaksi *Check-in Service* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Check-in Service	Sama Puas	0,657142857	0,328571429	0,014285714			1
	Puas	0,117647059	0,843137255	0,039215686			1
	Netral		0,727272727	0,272727273			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.22 probabilitas dari interaksi antara atribut *Check-in Service* dengan klasifikasi. Pada saat *Check-in Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,66 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.22 Probabilitas Interaksi *Self Service Check-in* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Self Service Check-in	Sama Puas	0,745454545	0,254545455				1
	Puas	0,185185185	0,796296296	0,018518519			1
	Netral	0,045454545	0,727272727	0,227272727			1
	Tidak Puas						0
	Sama Sekali Tidak Puas		0,5			0,5	1

Terlihat pada tabel 5.23 probabilitas dari interaksi antara atribut *Customer Service Service* klasifikasi. Pada saat *Customer Service Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,77 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.23 Probabilitas Interaksi *Customer Service Service* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Costomer Service	Sama Puas	0,770491803	0,229508197				1
	Puas	0,086206897	0,844827586	0,068965517			1
	Netral		0,846153846	0,153846154			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.24 probabilitas dari interaksi antara atribut *Executive Lounge* klasifikasi. Pada saat *Executive Lounge* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,77 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.24 Probabilitas Interaksi *Executive Lounge* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Executive Lounge	Sama Puas	0,775862069	0,224137931				1
	Puas	0,1	0,86	0,04			1
	Netral	0,086956522	0,782608696	0,130434783			1
	Tidak Puas						0
	Sama Sekali Tidak Puas			0,5		0,5	1

Terlihat pada tabel 5.25 probabilitas dari interaksi antara atribut *Boarding Management* klasifikasi. Pada saat *Boarding Management* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,79 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.25 Probabilitas Interaksi *Boarding Management* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Boarding Management	Sama Puas	0,796610169	0,203389831				1
	Puas	0,090909091	0,872727273	0,036363636			1
	Netral		0,785714286	0,214285714			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas		0,333333333	0,333333333		0,333333333	1

Terlihat pada tabel 5.26 probabilitas dari interaksi antara atribut *On Time Perfomance* klasifikasi. Pada saat *On Time Perfomance* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,77 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.26 Probabilitas Interaksi *On Time Perfomance* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
On Time Perfomance	Sama Puas	0,793103448	0,206896552				1
	Puas	0,105263158	0,859649123	0,035087719			1
	Netral		0,8125	0,1875			1
	Tidak Puas						0
	Sama Sekali Tidak Puas			0,5		0,5	1

Terlihat pada tabel 5.27 probabilitas dari interaksi antara atribut *Cabin Crew Service* klasifikasi. Pada saat *Cabin Crew Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,59 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.27 Probabilitas Interaksi *Cabin Crew Service* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Crew Service	Sama Puas	0,559139785	0,408602151	0,032258065			1
	Puas		0,947368421	0,052631579			1
	Netral			1			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.28 probabilitas dari interaksi antara atribut *Cabin Condition* klasifikasi. Pada saat *Cabin Condition* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,62 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.28 Probabilitas Interaksi *Cabin Condition* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Condition	Sama Puas	0,62962963	0,358024691	0,012345679			1
	Puas	0,021276596	0,893617021	0,085106383			1
	Netral		0,75	0,25			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.29 probabilitas dari interaksi antara atribut *Seat Comfort* klasifikasi. Pada saat *Seat Comfort* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,71 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.29 Probabilitas Interaksi *Seat Comfort* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Seat Comfort	Sama Puas	0,718309859	0,281690141				1
	Puas	0,018181818	0,909090909	0,072727273			1
	Netral		0,666666667	0,333333333			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.30 probabilitas dari interaksi antara atribut *Lavatory* klasifikasi. Pada saat *Lavatory* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,78 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.30 Probabilitas Interaksi *Lavatory* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Lavatory	Sama Puas	0,78125	0,21875				1
	Puas	0,031746032	0,904761905	0,063492063			1
	Netral		0,6	0,4			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.31 probabilitas dari interaksi antara atribut *Food & Beverage* klasifikasi. Pada saat *Food & Beverage* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,73 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.31 Probabilitas Interaksi *Food & Beverage* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Food & Beverage	Sama Puas	0,734375	0,265625				1
	Puas	0,092592593	0,87037037	0,037037037			1
	Netral		0,714285714	0,285714286			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.32 probabilitas dari interaksi antara atribut *In Flight Entertainment* klasifikasi. Pada saat *In Flight Entertainment* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,79 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.32 Probabilitas Interaksi *In Flight Entertainment* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
In Flight Entertainment	Sama Puas	0,790322581	0,209677419				1
	Puas	0,054545455	0,927272727	0,018181818			1
	Netral		0,642857143	0,357142857			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.33 probabilitas dari interaksi antara atribut *Reading Material* klasifikasi. Pada saat *Reading Material* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,75 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.33 Probabilitas Interaksi *Reading Material* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Reading Material	Sama Puas	0,754385965	0,245614035				1
	Puas	0,126984127	0,825396825	0,047619048			1
	Netral	0,083333333	0,666666667	0,25			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.34 probabilitas dari interaksi antara atribut *Cabin Amenity* klasifikasi. Pada saat *Cabin Amenity* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,79 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.34 Probabilitas Interaksi *Cabin Amenity* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Amenity	Sama Puas	0,793103448	0,206896552				1
	Puas	0,081967213	0,885245902	0,032786885			1
	Netral	0,076923077	0,615384615	0,307692308			1
	Tidak Puas						0
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.35 probabilitas dari interaksi antara atribut *In Flight Sales* klasifikasi. Pada saat *In Flight Sales* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,8 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.35 Probabilitas Interaksi *In Flight Sales* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
In Flight Sales	Sama Puas	0,807692308	0,192307692				1
	Puas	0,195652174	0,782608696	0,02173913			1
	Netral	0,033333333	0,8	0,166666667			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.36 probabilitas dari interaksi antara atribut *Baggage* klasifikasi. Pada saat *Baggage* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,73 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.36 Probabilitas Interaksi *Baggage* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Baggage	Sama Puas	0,736842105	0,263157895				1
	Puas	0,14516129	0,806451613	0,048387097			1
	Netral	0,090909091	0,636363636	0,272727273			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.37 probabilitas dari interaksi antara atribut *Garuda Frequent Flyer* klasifikasi. Pada saat *Garuda Frequent Flyer* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,75 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.37 Probabilitas Interaksi *Garuda Frequent Flyer* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Garuda Frequent Flyer	Sama Puas	0,754385965	0,245614035				1
	Puas	0,129032258	0,838709677	0,032258065			1
	Netral	0,083333333	0,25	0,666666667			1
	Tidak Puas			1			1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.38 probabilitas dari interaksi antara atribut *Delay Management* klasifikasi. Pada saat *Delay Management* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,77 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.38 Probabilitas Interaksi *Delay Management* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Delay Management	Sama Puas	0,775510204	0,224489796				1
	Puas	0,269230769	0,711538462	0,019230769			1
	Netral		0,857142857	0,142857143			1
	Tidak Puas		0,666666667	0,333333333			1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.39 probabilitas dari interaksi antara atribut *On Service Recovery* klasifikasi. Pada saat *Service Recovery* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,75 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.39 Probabilitas Interaksi *Service Recovery* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Service Recovery	Sama Puas	0,75862069	0,24137931				1
	Puas	0,125	0,839285714	0,035714286			1
	Netral	0,066666667	0,266666667	0,666666667			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas					1	1

Terlihat pada tabel 5.40 probabilitas dari interaksi antara *Class* penerbangan yang digunakan koresponden dengan klasifikasi. Pada saat koresponden berada *Class* penerbangan bisnis dan klasifikasi sangat puas memiliki probabilitas 0,66 pada kelas penerbangan bisnis. Probabilitas pada kelas penerbangan yang responden gunakan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.40 Probabilitas Interaksi *Class* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Class	Bisnis	0,666666667	0,277777778			0,055555556	1
	Economy	0,350877193	0,596491228	0,052631579			1
	Tidak Di Isi		1				1

Terlihat pada tabel 5.41 probabilitas dari interaksi antara *Gender dengan klasifikasi*. Pada saat *gender* responden pria dan klasifikasi sangat puas memiliki probabilitas 0,45 pada tingkat kepuasan sangat puas. Probabilitas gender responden dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.41 Probabilitas Interaksi *Gender* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Gender	Pria	0,459459459	0,513513514	0,013513514		0,013513514	1
	Wanita	0,2	0,65	0,15			1
	Tidak Di Isi	0,358974359	0,58974359	0,051282051			1

Terlihat pada tabel 5.42 probabilitas dari interaksi usia responden dengan klasifikasi. Pada saat responden memiliki usia 20—29 tahun memiliki probabilitas tinggi pada klasifikasi sangat puas. Probabilitas pada usia responden dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.42 Probabilitas Interaksi *Age* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Age	Dibawah 20		1				1
	20 - 29	0,5	0,447368421	0,052631579			1
	30 - 39	0,347826087	0,652173913				1
	40 - 49	0,428571429	0,457142857	0,085714286		0,028571429	1
	50 - 59	0,25	0,75				1
	Diatas 60	0,333333333	0,333333333	0,333333333			1
	Tidak Di Isi	0,4	0,6				1

Terlihat pada tabel 5.43 probabilitas dari interaksi antara tujuan penerbangan responden dengan klasifikasi. Pada saat respoden memiliki tujuan penerbangan bisnis memiliki probabilitas tinggi pada klasifikasi tingkat kepuasan puas. Probabilitas pada tujuan penerbangan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.44 Probabilitas Interaksi *Purpose* Jakarta – Surabaya

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Purpose	Bisnis	0,471428571	0,5	0,028571429			1
	Liburan	0,25	0,5	0,125		0,125	1
	Tujuan Pribadi	0,379310345	0,551724138	0,068965517			1
	Seminar	0,272727273	0,727272727				1
	Lain-lain		0,8	0,2			1
	Tidak Di Isi	0,3	0,7				1

Setelah didapat probabilitas interaksi dari tiap atribut dengan klasifikasi, tahap selanjutnya dalam penggunaan klasifikasi dengan

metode bayes adalah menghitung P (At|K) atau dikenal dengan probabilitas interaksi atribut dengan klasifikasi. P (At|K) didapat dengan cara penggalian tiap probabilitas interaksi dengan klasifikasi pada responden.

$$P(At \mid K) = P(At|K)_1 \; xP(At|K)_2 \; x \ldots x \, P(At|K)_n \quad \ldots\ldots\ldots\ldots (5.6)$$

Setelah itu dilakukan perhitungan P(K) P(At|K). Perhitungan tersebut dengan cara penggalian dari probabilitas prior dari klasifikasi dengan probabilitas interaksi yang sudah didapat.

Selanjutnya menghitung probabilitas posterior, probabilitas posterior bisa didapat dengan cara sebagai berikut :

$$\text{Probabilitas Posterior} = \frac{(P(K)\ P(At|K))}{(P\ (T))} \quad \ldots\ldots\ldots\ldots\ldots (5.7)$$

Hasil dari perhitungan probabilitas posterior pada penerbangan Jakarta – Surabaya dengan menggunakan metode bayes dapat dilihat pada tabel 5.45.

Tabel 5.45 Hasil Perhitungan Probabilitas Posterior Penerbangan Jakarta – Surabaya

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
1	7,41047E-05	Sangat Puas	0,390977444	2,89733E-05	0,609022556	4,75734E-05
2	4,55029E-05	Sangat Puas	0,390977444	1,77906E-05	0,609022556	2,92117E-05
3	4,55029E-05	Sangat Puas	0,390977444	1,77906E-05	0,609022556	2,92117E-05
4	4,55029E-05	Sangat Puas	0,390977444	1,77906E-05	0,609022556	2,92117E-05
5	2,96506E-08	Sangat Puas	0,390977444	1,15927E-08	0,609022556	1,90349E-08
6	6,94314E-08	Sangat Puas	0,390977444	2,71461E-08	0,609022556	4,45732E-08
7	5,46535E-11	Sangat Puas	0,390977444	2,13683E-11	0,609022556	3,50862E-11
8	6,71124E-06	Puas	0,556390977	3,73407E-06	0,443609023	8,41749E-06
9	1,03741E-10	Sangat Puas	0,390977444	4,05605E-11	0,609022556	6,65993E-11
10	4,89431E-05	Puas	0,556390977	2,72315E-05	0,443609023	6,13863E-05
11	2,12295E-07	Puas	0,556390977	1,18119E-07	0,443609023	2,66268E-07
12	0,001425097	Puas	0,556390977	0,000792911	0,443609023	0,001787409
13	0,000155109	Puas	0,556390977	8,6301E-05	0,443609023	0,000194543
14	2,92284E-10	Puas	0,556390977	1,62624E-10	0,443609023	3,66593E-10
15	0,000710548	Puas	0,556390977	0,000395342	0,443609023	0,000891195
16	0,000804337	Puas	0,556390977	0,000447526	0,443609023	0,001008829
17	0,001583018	Puas	0,556390977	0,000880777	0,443609023	0,001985481
18	2,56787E-30	Netral	0,045112782	1,15844E-31	0,954887218	1,21317E-31
19	0,004507978	Puas	0,556390977	0,002508198	0,443609023	0,005654075
20	2,26131E-05	Puas	0,556390977	1,25817E-05	0,443609023	2,83622E-05
21	0,003598937	Puas	0,556390977	0,002002416	0,443609023	0,004513921
22	0,003677569	Puas	0,556390977	0,002046166	0,443609023	0,004612545
23	1,04897E-27	Netral	0,045112782	4,73218E-29	0,954887218	4,95575E-29
24	0,006033512	Puas	0,556390977	0,003356992	0,443609023	0,007567456
25	0,000340573	Puas	0,556390977	0,000189492	0,443609023	0,000427159
26	1,44155E-07	Puas	0,556390977	8,02064E-08	0,443609023	1,80804E-07
27	1,3502E-07	Puas	0,556390977	7,51239E-08	0,443609023	1,69347E-07
28	1,96832E-07	Puas	0,556390977	1,09516E-07	0,443609023	2,46874E-07
29	3,23479E-05	Puas	0,556390977	1,79981E-05	0,443609023	4,05719E-05
30	0,000137416	Puas	0,556390977	7,64571E-05	0,443609023	0,000172352
31	2,89564E-05	Sangat Puas	0,390977444	1,13213E-05	0,609022556	1,85893E-05
32	9,48483E-05	Sangat Puas	0,390977444	3,70836E-05	0,609022556	6,08903E-05
33	9,48483E-05	Sangat Puas	0,390977444	3,70836E-05	0,609022556	6,08903E-05
34	8,90339E-05	Sangat Puas	0,390977444	3,48102E-05	0,609022556	5,71575E-05
35	1,4069E-09	Sangat Puas	0,390977444	5,50066E-10	0,609022556	9,03194E-10

Tabel 5.45 Hasil Perhitungan Probabilitas Posterior Penerbangan Jakarta – Surabaya (Lanjutan)

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
36	4,67871E-09	Sangat Puas	0,390977444	1,82927E-09	0,609022556	3,00362E-09
37	4,28918E-13	Sangat Puas	0,390977444	1,67697E-13	0,609022556	2,75355E-13
38	7,8644E-11	Sangat Puas	0,390977444	3,0748E-11	0,609022556	5,04875E-11
39	1,56272E-26	Netral	0,045112782	7,04985E-28	0,954887218	7,38291E-28
40	1,29788E-05	Puas	0,556390977	7,22128E-06	0,443609023	1,62785E-05
41	3,55406E-05	Puas	0,556390977	1,97745E-05	0,443609023	4,45764E-05
42	1,99651E-05	Puas	0,556390977	1,11084E-05	0,443609023	2,5041E-05
43	0,00289855	Puas	0,556390977	0,001612727	0,443609023	0,00363547
44	0,002010547	Puas	0,556390977	0,00111865	0,443609023	0,002521703
45	0,001446995	Puas	0,556390977	0,000805095	0,443609023	0,001814875
46	0,000167612	Puas	0,556390977	9,32579E-05	0,443609023	0,000210225
47	8,61528E-18	Netral	0,045112782	3,88659E-19	0,954887218	4,07021E-19
48	3,16917E-08	Puas	0,556390977	1,7633E-08	0,443609023	3,97489E-08
49	5,53153E-06	Puas	0,556390977	3,0777E-06	0,443609023	6,93786E-06
50	4,98214E-11	Puas	0,556390977	2,77202E-11	0,443609023	6,24878E-11
51	1,67456E-06	Puas	0,556390977	9,31712E-07	0,443609023	2,1003E-06
52	9,15089E-10	Sangat Puas	0,390977444	3,57779E-10	0,609022556	5,87465E-10
53	1,12502E-06	Puas	0,556390977	6,25952E-07	0,443609023	1,41104E-06
54	2,21074E-05	Puas	0,556390977	1,23004E-05	0,443609023	2,7728E-05
55	0,001864666	Puas	0,556390977	0,001037483	0,443609023	0,002338734
56	0,000974524	Puas	0,556390977	0,000542216	0,443609023	0,001222284
57	1,1499E-23	Netral	0,045112782	5,18752E-25	0,954887218	5,4326E-25
58	2,64714E-11	Sangat Puas	0,390977444	1,03497E-11	0,609022556	1,6994E-11
59	1,28372E-07	Sangat Puas	0,390977444	5,01907E-08	0,609022556	8,24119E-08
60	4,05149E-05	Sangat Puas	0,390977444	1,58404E-05	0,609022556	2,60096E-05
61	7,50789E-10	Puas	0,556390977	4,17732E-10	0,443609023	9,41667E-10
62	1,1372E-12	Sangat Puas	0,390977444	4,44621E-13	0,609022556	7,30056E-13
63	0,003605132	Puas	0,556390977	0,002005863	0,443609023	0,004521691
64	2,00971E-07	Puas	0,556390977	1,11819E-07	0,443609023	2,52066E-07
65	5,81048E-07	Puas	0,556390977	3,2329E-07	0,443609023	7,28773E-07
66	3,71369E-05	Puas	0,556390977	2,06626E-05	0,443609023	4,65785E-05
67	2,2152E-08	Puas	0,556390977	1,23252E-08	0,443609023	2,77838E-08
68	3,16542E-05	Sangat Puas	0,390977444	1,23761E-05	0,609022556	2,03212E-05
69	1,11203E-05	Puas	0,556390977	6,18725E-06	0,443609023	1,39475E-05
70	4,6098E-09	Puas	0,556390977	2,56485E-09	0,443609023	5,78179E-09
71	5,82402E-05	Sangat Puas	0,390977444	2,27706E-05	0,609022556	3,73888E-05
72	0,00200138	Puas	0,556390977	0,00111355	0,443609023	0,002510205
73	1,20925E-05	Sangat Puas	0,390977444	4,72788E-06	0,609022556	7,76306E-06
74	5,80174E-08	Puas	0,556390977	3,22803E-08	0,443609023	7,27675E-08
75	3,74879E-05	Sangat Puas	0,390977444	1,46569E-05	0,609022556	2,40663E-05
76	2,42314E-07	Puas	0,556390977	1,34821E-07	0,443609023	3,03919E-07
77	0,000142335	Puas	0,556390977	7,91938E-05	0,443609023	0,000178522
78	2,343E-05	Sangat Puas	0,390977444	9,16059E-06	0,609022556	1,50415E-05
79	0,00158418	Puas	0,556390977	0,000881424	0,443609023	0,001986938
80	1,83058E-05	Sangat Puas	0,390977444	7,15714E-06	0,609022556	1,17518E-05
81	1,81781E-10	Sangat Puas	0,390977444	7,10721E-11	0,609022556	1,16699E-10
82	2,28176E-13	Sangat Puas	0,390977444	8,92117E-14	0,609022556	1,46483E-13
83	4,88349E-05	Puas	0,556390977	2,71713E-05	0,443609023	6,12505E-05
84	1,29102E-06	Puas	0,556390977	7,18309E-07	0,443609023	1,61924E-06
85	6,78618E-06	Puas	0,556390977	3,77577E-06	0,443609023	8,51148E-06
86	3,66115E-05	Sangat Puas	0,390977444	1,43143E-05	0,609022556	2,35037E-05
87	9,48483E-05	Sangat Puas	0,390977444	3,70836E-05	0,609022556	6,08903E-05
88	2,74657E-13	Sangat Puas	0,390977444	1,07385E-13	0,609022556	1,76323E-13
89	5,96954E-08	Puas	0,556390977	3,3214E-08	0,443609023	7,48722E-08

Tabel 5.45 Hasil Perhitungan Probabilitas Posterior Penerbangan Jakarta – Surabaya (Lanjutan)

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
90	0,003211174	Puas	0,556390977	0,001786668	0,443609023	0,004027574
91	2,61404E-08	Sangat Puas	0,390977444	1,02203E-08	0,609022556	1,67815E-08
92	2,88793E-09	Puas	0,556390977	1,60682E-09	0,443609023	3,62215E-09
93	3,61655E-06	Sangat Puas	0,390977444	1,41399E-06	0,609022556	2,32174E-06
94	5,90413E-08	Puas	0,556390977	3,285E-08	0,443609023	7,40518E-08
95	8,17901E-10	Puas	0,556390977	4,55073E-10	0,443609023	1,02584E-09
96	1,35977E-07	Puas	0,556390977	7,56563E-08	0,443609023	1,70547E-07
97	2,53516E-05	Sangat Puas	0,390977444	9,91191E-06	0,609022556	1,62751E-05
98	2,00637E-26	Netral	0,045112782	9,05127E-28	0,954887218	9,47889E-28
99	6,32237E-05	Puas	0,556390977	3,51771E-05	0,443609023	7,92976E-05
100	4,15514E-08	Puas	0,556390977	2,31188E-08	0,443609023	5,21153E-08
101	7,03071E-11	Sangat Puas	0,390977444	2,74885E-11	0,609022556	4,51354E-11
102	4,97484E-09	Sangat Puas	0,390977444	1,94505E-09	0,609022556	3,19372E-09
103	2,04612E-08	Puas	0,556390977	1,13844E-08	0,443609023	2,56632E-08
104	0,000852778	Puas	0,556390977	0,000474478	0,443609023	0,001069586
105	1,66962E-08	Puas	0,556390977	9,28962E-09	0,443609023	2,0941E-08
106	2,69957E-08	Puas	0,556390977	1,50202E-08	0,443609023	3,3859E-08
107	1,11719E-07	Sama Sekali Tidak Puas	0,007518797	8,39991E-10	0,992481203	8,46355E-10
108	1,26761E-09	Sangat Puas	0,390977444	4,95608E-10	0,609022556	8,13776E-10
109	2,27514E-05	Sangat Puas	0,390977444	8,8953E-06	0,609022556	1,46059E-05
110	0,001784534	Puas	0,556390977	0,000992899	0,443609023	0,002238229
111	0,000110656	Sangat Puas	0,390977444	4,32641E-05	0,609022556	7,10387E-05
112	1,9201E-09	Puas	0,556390977	1,06833E-09	0,443609023	2,40826E-09
113	1,18912E-07	Sangat Puas	0,390977444	4,64921E-08	0,609022556	7,63388E-08
114	0,000167152	Puas	0,556390977	9,30016E-05	0,443609023	0,000209648
115	5,82402E-05	Sangat Puas	0,390977444	2,27706E-05	0,609022556	3,73888E-05
116	4,05149E-05	Sangat Puas	0,390977444	1,58404E-05	0,609022556	2,60096E-05
117	0,001256761	Puas	0,556390977	0,00069925	0,443609023	0,001576276
118	2,91201E-05	Sangat Puas	0,390977444	1,13853E-05	0,609022556	1,86944E-05
119	4,68599E-05	Sangat Puas	0,390977444	1,83212E-05	0,609022556	3,00829E-05
120	2,88794E-05	Sangat Puas	0,390977444	1,12912E-05	0,609022556	1,85399E-05
121	3,87559E-05	Sangat Puas	0,390977444	1,51527E-05	0,609022556	2,48803E-05
122	0,00664998	Puas	0,556390977	0,003699989	0,443609023	0,008340653
123	0,000110656	Sangat Puas	0,390977444	4,32641E-05	0,609022556	7,10387E-05
124	2,27427E-08	Puas	0,556390977	1,26539E-08	0,443609023	2,85248E-08
125	2,73161E-05	Puas	0,556390977	1,51984E-05	0,443609023	3,42608E-05
126	2,04765E-12	Sangat Puas	0,390977444	8,00584E-13	0,609022556	1,31454E-12
127	2,06583E-08	Puas	0,556390977	1,14941E-08	0,443609023	2,59104E-08
128	2,29192E-07	Puas	0,556390977	1,2752E-07	0,443609023	2,87461E-07
129	2,88794E-05	Sangat Puas	0,390977444	1,12912E-05	0,609022556	1,85399E-05
130	1,45853E-10	Puas	0,556390977	8,11511E-11	0,443609023	1,82934E-10
131	0,000667031	Puas	0,556390977	0,00037113	0,443609023	0,000836615
132	1,74839E-05	Sangat Puas	0,390977444	6,8358E-06	0,609022556	1,12242E-05
133	1,01989E-05	Sangat Puas	0,390977444	3,98755E-06	0,609022556	6,54746E-06

Berdasarkan hasil perhitungan diatas dapat dibuat rangkuman yang dapat dilihat pada tabel 5.46.

Tabel 5.46 Rangkuman Probabilitas Posterior Penerbangan Jakarta – Surabaya

Klasifikasi	Probabilitas Prior	Probabilitas Posterior
Sangat Puas	0,390977444	0,000926908
Puas	0,556390977	0,068565567
Netral	0,045112782	4,07021E-19
Tidak Puas	0	0
Sama Sekali Tidak Puas	0,007518797	8,46355E-10

Probabilitas prior adalah probabilitas yang didapatkan dari data awal penelitian, sedangkan probabilitas posterior adalah Probabilitas yang telah diperbaiki setelah mendapat informasi tambahan atau setelah melakukan penelitian. Penarikan kesimpulan pada metode bayes menggunakan hasil data probabilitas posterior dengan cara melihat hasil dari probabilitas posterior. Kesimpulan diambil pada saat nilai probabilitas posterior tertinggi pada klasifikasi atau dapat di rumuskan sebagai berikut :

h_{MAP} = arg max P(x | h) P(h) .....(5,8)

Pada penelitian ini penerbangan Jakarta – Surabaya berdasarkan data tabel 5.68 dapat ditarik kesimpulan bahwa tingkat kepuasan pelayanan yang diberikan perusahaan kepada responden berada pada tingkat kepuasan puas. Hasil dari ini memiliki kesamaan dengan hasil *data mining* dengan menggunakan *software* weka sehingga kesimpulan dapat diterima.

2. Penerbangan Surabaya – Jakarta

Tahap awal dalam perhitungan manual data mining menggunakan metode bayes adalah dengan cara menentukan probabilitas posterior klasifikasi terlebih dahulu. Probabilitas prior didapat dengan cara pembagian jumlah salah satu klasifikasi dengan jumlah koresponden, atau dapat dilihat pada rumus 5.9.

Probabilitas Prior = $\frac{x_i}{\sum_{i=1}^{n} x}$ .....(5.9)

Pada penerbangan Jakarta – Surabaya memiliki probabilitas posterior yang memiliki nilai besar pada klasifikasi puas dengan nilai diata 0,5. Data hasil probabilitas posterior selengkapnya dapat dilihat pada tabel 5.47.

Tabel 5.47 Probabilitas Prior Penerbangan Surabaya – Jakarta

Klasifikasi	Count	Probabilitas Prior
Sama Puas	53	0,365517241
Puas	88	0,606896552
Netral	3	0,020689655
Tidak Puas	1	0,006896552
Sama Sekali Tidak Puas	0	0
Total	145	1

Tahap selanjutnya setelah mendapatkan probabilitas prior adalah mendapatkan probabilitas interaksi antara tiap atribut dengan klasifikasi. Probabilitas interaksi atribut dengan klasifikasi dapat dilihat pada tabel berikutnya.

Terlihat pada tabel 5.48 probabilitas dari interaksi antara atribut *Website Service* dengan klasifikasi. Pada saat *Website Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,71 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.48 Probabilitas Interaksi *Website Service* Penerbangan Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Website Service	Sama Puas	0,71641791	0,28358209				1
	Puas	0,081967213	0,901639344	0,016393443			1
	Netral		0,8	0,133333333	0,066666667		1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.49 probabilitas dari interaksi antara atribut *Contact Center Service* dengan klasifikasi. Pada saat *Contact Center Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,74 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.49 Probabilitas Interaksi *Contact Center Service* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Contact Center Service	Sama Puas	0,74137931	0,25862069				1
	Puas	0,14084507	0,830985915	0,028169014			1
	Netral		0,933333333	0,066666667			1
	Tidak Puas				1		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.50 probabilitas dari kemungkinan antara atribut *Sales Office Service* dengan klasifikasi. Pada saat *Sales Office Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,70 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.50 Probabilitas Interaksi *Sales Office Service* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Sales Office Service	Sama Puas	0,707692308	0,292307692				1
	Puas	0,107692308	0,861538462	0,030769231			1
	Netral		0,928571429	0,071428571			1
	Tidak Puas						0
	Sama Sekali Tidak Puas				1		1

Terlihat pada tabel 5.51 probabilitas dari interaksi antara atribut *Check-in Service* dengan klasifikasi. Pada saat *Check-in Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,59 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.51 Probabilitas Interaksi *Check-in Service* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Check-in Service	Sama Puas	0,595238095	0,404761905				1
	Puas	0,056603774	0,905660377	0,037735849			1
	Netral		0,833333333	0,166666667			1
	Tidak Puas		0,5		0,5		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.52 probabilitas dari interaksi antara atribut *Check-in Service* dengan klasifikasi. Pada saat *Check-in Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,71 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.52 Probabilitas Interaksi *Self Service Check-in* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Self Service Check-in	Sama Puas	0,712121212	0,287878788				1
	Puas	0,095238095	0,904761905				1
	Netral		0,857142857	0,142857143			1
	Tidak Puas			0,5	0,5		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.53 probabilitas dari interaksi antara atribut *Customer Service Service* klasifikasi. Pada saat *Customer Service Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,68 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.53 Probabilitas Interaksi *Customer Service Service* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Costomer Service	Sama Puas	0,686567164	0,313432836				1
	Puas	0,106060606	0,863636364	0,03030303			1
	Netral		1				1
	Tidak Puas			1			1
	Sama Sekali Tidak Puas				1		1

Terlihat pada tabel 5.54 probabilitas dari interaksi antara atribut *Executive Lounge* klasifikasi. Pada saat *Executive Lounge* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,64 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.54 Probabilitas Interaksi *Executive Lounge* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Executive Lounge	Sama Puas	0,640625	0,359375				1
	Puas	0,2	0,766666667	0,033333333			1
	Netral		0,904761905	0,047619048	0,047619048		1
	Tidak Puas						0
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.55 probabilitas dari interaksi antara atribut *Boarding Management* klasifikasi. Pada saat *Boarding Management* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,73 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.55 Probabilitas Interaksi *Boarding Management* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Boarding Management	Sama Puas	0,738461538	0,261538462				1
	Puas	0,0625	0,9375				1
	Netral	0,071428571	0,714285714	0,214285714			1
	Tidak Puas		0,5		0,5		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.56 probabilitas dari interaksi antara atribut *On Time Perfomance* klasifikasi. Pada saat *On Time Perfomance* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,60 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.56 Probabilitas Interaksi *On Time Perfomance* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
On Time Perfomance	Sama Puas	0,605263158	0,394736842				1
	Puas	0,12	0,86	0,02			1
	Netral	0,066666667	0,933333333				1
	Tidak Puas		0,333333333	0,333333333	0,333333333		1
	Sama Sekali Tidak Puas			1			1

Terlihat pada tabel 5.57 probabilitas dari interaksi antara atribut *Cabin Crew Service* klasifikasi. Pada saat *Cabin Crew Service* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,47 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.57 Probabilitas Interaksi *Cabin Crew Service* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Crew Service	Sama Puas	0,47706422	0,513761468	0,009174312			1
	Puas	0,028571429	0,914285714	0,057142857			1
	Netral						0
	Tidak Puas				1		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.58 probabilitas dari interaksi antara atribut *Cabin Condition* klasifikasi. Pada saat *Cabin Condition* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,52 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.58 Probabilitas Interaksi *Cabin Condition* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Condition	Sama Puas	0,528735632	0,471264368				1
	Puas	0,125	0,839285714	0,035714286			1
	Netral			1			1
	Tidak Puas				1		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.59 probabilitas dari interaksi antara atribut *Seat Comfort* klasifikasi. Pada saat *Seat Comfort* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,64 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.59 Probabilitas Interaksi *Seat Comfort* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Seat Comfort	Sama Puas	0,649350649	0,350649351				1
	Puas	0,048387097	0,903225806	0,048387097			1
	Netral		1				1
	Tidak Puas		0,5		0,5		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.60 probabilitas dari interaksi antara atribut *Lavatory* klasifikasi. Pada saat *Lavatory* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,60 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.60 Probabilitas Interaksi *Lavatory* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Lavatory	Sama Puas	0,604938272	0,395061728				1
	Puas	0,070175439	0,877192982	0,052631579			1
	Netral		1				1
	Tidak Puas				1		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.61 probabilitas dari interaksi antara atribut *Food & Beverage* klasifikasi. Pada saat *Food & Beverage* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,73 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.61 Probabilitas Interaksi *Food & Beverage* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Food & Beverage	Sama Puas	0,74137931	0,25862069				1
	Puas	0,121621622	0,824324324	0,040540541	0,013513514		1
	Netral	0,076923077	0,923076923				1
	Tidak Puas						0
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.62 probabilitas dari interaksi antara atribut *In Flight Entertainment* klasifikasi. Pada saat *In Flight Entertainment* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,79 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.62 Probabilitas Interaksi *In Flight Entertainment* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
In Flight Entertainment	Sama Puas	0,638888889	0,347222222	0,013888889			1
	Puas	0,107142857	0,892857143				1
	Netral	0,071428571	0,785714286	0,142857143			1
	Tidak Puas		0,666666667		0,333333333		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.63 probabilitas dari interaksi antara atribut *Reading Material* klasifikasi. Pada saat *Reading Material* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,65 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.63 Probabilitas Interaksi *Reading Material* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Reading Material	Sama Puas	0,65	0,333333333	0,016666667			1
	Puas	0,208955224	0,776119403	0,014925373			1
	Netral		0,933333333	0,066666667			1
	Tidak Puas		1				1
	Sama Sekali Tidak Puas		0,5		0,5		1

Terlihat pada tabel 5.64 probabilitas dari interaksi antara atribut *Cabin Amenity* klasifikasi. Pada saat *Cabin Amenity* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,72 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.64 Probabilitas Interaksi *Cabin Amenity* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Cabin Amenity	Sama Puas	0,721311475	0,278688525				1
	Puas	0,140625	0,84375	0,015625			1
	Netral		0,9375	0,0625			1
	Tidak Puas		0,5	0,25	0,25		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.65 probabilitas dari interaksi antara atribut *In Flight Sales* klasifikasi. Pada saat *In Flight Sales* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,7 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.65 Probabilitas Interaksi *In Flight Sales* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
In Flight Sales	Sama Puas	0,7	0,3				1
	Puas	0,156862745	0,823529412	0,019607843			1
	Netral	0,12	0,84	0,04			1
	Tidak Puas		0,75	0,25			1
	Sama Sekali Tidak Puas		1				1

Terlihat pada tabel 5.66 probabilitas dari interaksi antara atribut *Baggage* klasifikasi. Pada saat *Baggage* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,7 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.66 Probabilitas Interaksi *Baggage* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Baggage	Sama Puas	0,7	0,3				1
	Puas	0,06779661	0,915254237	0,016949153			1
	Netral		0,777777778	0,222222222			1
	Tidak Puas		0,75		0,25		1
	Sama Sekali Tidak Puas		1				1

Terlihat pada tabel 5.67 probabilitas dari interaksi antara atribut *Garuda Frequent Flyer* klasifikasi. Pada saat *Garuda Frequent Flyer* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,65 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.67 Probabilitas Interaksi *Garuda Frequent Flyer* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Garuda Frequent Flyer	Sama Puas	0,652777778	0,347222222				1
	Puas	0,115384615	0,826923077	0,038461538	0,019230769		1
	Netral		1				1
	Tidak Puas		0,666666667	0,333333333			1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.68 probabilitas dari interaksi antara atribut *Delay Management* klasifikasi. Pada saat *Delay Management* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,77 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel di bawah.

Tabel 5.68 Probabilitas Interaksi *Delay Management* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Delay Management	Sama Puas	0,773584906	0,226415094				1
	Puas	0,147058824	0,838235294	0,014705882			1
	Netral	0,052631579	0,947368421				1
	Tidak Puas	0,2	0,2	0,4	0,2		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.69 probabilitas dari interaksi antara atribut *On Service Recovery* klasifikasi. Pada saat *Service Recovery* berada tingkat kepuasan sangat puas dan klasifikasi sangat puas memiliki probabilitas 0,71 pada tingkat kepuasan sangat puas. Probabilitas pada tingkat kepuasan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.69 Probabilitas Interaksi *Service Recovery* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Service Recovery	Sama Puas	0,714285714	0,285714286				1
	Puas	0,1	0,883333333	0,016666667			1
	Netral	0,105263158	0,842105263	0,052631579			1
	Tidak Puas		0,333333333	0,333333333	0,333333333		1
	Sama Sekali Tidak Puas						0

Terlihat pada tabel 5.70 probabilitas dari interaksi antara *Class* penerbangan yang digunakan koresponden dengan klasifikasi. Pada saat koresponden berada *Class* penerbangan bisnis dan klasifikasi sangat puas memiliki probabilitas 0,28 pada kelas penerbangan bisnis. Probabilitas pada kelas penerbangan yang responden gunakan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.70 Probabilitas Interaksi *Class* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Class	Bisnis	0,285714286	0,666666667		0,047619048		1
	Economy	0,37704918	0,598360656	0,024590164			1
	Tidak Di Isi	0,5	0,5				1

Terlihat pada tabel 5.71 probabilitas dari interaksi antara *Gender dengan klasifikasi*. Pada saat *gender* responden pria dan klasifikasi sangat puas memiliki probabilitas 0,32 pada tingkat kepuasan sangat puas. Probabilitas gender responden dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.71 Probabilitas Interaksi *Gender* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Gender	Pria	0,329896907	0,659793814	0,010309278			1
	Wanita	0,6	0,35		0,05		1
	Tidak Di Isi	0,321428571	0,607142857	0,071428571			1

Terlihat pada tabel 5.72 probabilitas dari interaksi usia responden dengan klasifikasi. Pada saat responden memiliki usia 20—29 tahun memiliki probabilitas tinggi pada klasifikasi sangat puas. Probabilitas pada usia responden dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.72 Probabilitas Interaksi *Age* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Age	Dibawah 20	0,333333333	0,666666667				1
	20 - 29	0,451612903	0,516129032		0,032258065		1
	30 - 39	0,262295082	0,704918033	0,032786885			1
	40 - 49	0,333333333	0,666666667				1
	50 - 59	0,571428571	0,428571429				1
	Diatas 60	0,4	0,4	0,2			1
	Tidak Di Isi	0,5	0,5				1

Terlihat pada tabel 5.73 probabilitas dari interaksi antara tujuan penerbangan responden dengan klasifikasi. Pada saat respoden memiliki tujuan penerbangan bisnis memiliki probabilitas tinggi pada klasifikasi tingkat kepuasan puas. Probabilitas pada tujuan penerbangan dan klasifikasi lain dapat dilihat pada tabel dibawah.

Tabel 5.73 Probabilitas Interaksi *Purpose* Surabaya – Jakarta

	Klasifikasi	Sama Puas	Puas	Netral	Tidak Puas	Sama Sekali Tidak Puas	Total
Purpose	Bisnis	0,325	0,625	0,0375	0,0125		1
	Liburan	0,357142857	0,642857143			0,125	1,125
	Tujuan Pribadi	0,407407407	0,592592593				1
	Seminar	0,333333333	0,666666667				1
	Lain-lain	0,4	0,6				1
	Tidak Di Isi	0,714285714	0,285714286				1

Setelah didapat probabilitas interaksi dari tiap atribut dengan klasifikasi, tahap selanjutnya dalam penggunaan klasifikasi dengan metode bayes adalah menghitung P (At|K) atau dikenal dengan probabilitas interaksi atribut dengan klasifikasi. P (At|K) didapat dengan cara pengalian tiap probabilitas interaksi dengan klasifikasi pada responden.

$$P(At \mid K) = P(At|K)_1 \times P(At|K)_2 \times \ldots \times P(At|K)_n \quad \text{.....(5.10)}$$

Setelah itu dilakukan perhitungan P(K) P(At|K). Perhitungan tersebut dengan cara pengalian dari probabilitas prior dari klasifikasi dengan probabilitas interaksi yang sudah didapat.

Selanjutnya menghitung probabilitas posterior, probabilitas posterior bisa didapat dengan cara sebagai berikut :

$$\text{Probabilitas Posterior} = \frac{(P(K)\ P(At|K))}{(P\ (T))} \quad \text{.....(5.11)}$$

Hasil dari perhitungan probabilitas posterior pada penerbangan Surabaya – Jakarta dengan menggunakan metode bayes dapat dilihat pada tabel 5.96.

Tabel 5.74 Hasil Perhitungan Probabilitas Posterior Penerbangan Surabaya – Jakarta

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
1	6,99199E-06	Sangat Puas	0,365517241	2,55569E-06	0,634482759	4,02799E-06
2	1,81792E-06	Sangat Puas	0,365517241	6,6448E-07	0,634482759	1,04728E-06
3	6,20824E-08	Sangat Puas	0,365517241	2,26922E-08	0,634482759	3,57649E-08
4	2,06859E-05	Puas	0,606896552	1,25542E-05	0,393103448	3,19361E-05
5	2,35412E-08	Puas	0,606896552	1,42871E-08	0,393103448	3,63443E-08
6	3,98305E-06	Puas	0,606896552	2,4173E-06	0,393103448	6,14926E-06
7	5,85319E-06	Puas	0,606896552	3,55228E-06	0,393103448	9,03651E-06
8	0,001211259	Puas	0,606896552	0,000735109	0,393103448	0,001870014
9	7,05588E-05	Puas	0,606896552	4,28219E-05	0,393103448	0,000108933
10	5,12338E-37	Netral	0,020689655	1,06001E-38	0,979310345	1,0824E-38
11	1,13868E-07	Puas	0,606896552	6,9106E-08	0,393103448	1,75796E-07
12	0,000397048	Puas	0,606896552	0,000240967	0,393103448	0,000612986
13	0,008079692	Puas	0,606896552	0,004903537	0,393103448	0,01247391
14	3,64933E-08	Sangat Puas	0,365517241	1,33389E-08	0,634482759	2,10233E-08
15	0,000691292	Puas	0,606896552	0,000419543	0,393103448	0,001067258
16	4,10548E-06	Puas	0,606896552	2,4916E-06	0,393103448	6,33828E-06
17	0,00132589	Puas	0,606896552	0,000804678	0,393103448	0,002046988
18	1,59937E-06	Sangat Puas	0,365517241	5,84597E-07	0,634482759	9,21376E-07
19	1,59937E-06	Sangat Puas	0,365517241	5,84597E-07	0,634482759	9,21376E-07
20	9,77399E-08	Sangat Puas	0,365517241	3,57256E-08	0,634482759	5,63067E-08
21	1,24818E-09	Sangat Puas	0,365517241	4,56232E-10	0,634482759	7,19061E-10
22	3,27697E-05	Puas	0,606896552	1,98878E-05	0,393103448	5,05918E-05
23	1,29272E-30	Netral	0,020689655	2,67459E-32	0,979310345	2,7311E-32

Tabel 5.74 Hasil Perhitungan Probabilitas Posterior Penerbangan Surabaya – Jakarta (Lanjutan)

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
24	1,03407E-11	Sangat Puas	0,365517241	3,7797E-12	0,634482759	5,95714E-12
25	0,002273341	Puas	0,606896552	0,001379683	0,393103448	0,003509719
26	3,48338E-07	Puas	0,606896552	2,11405E-07	0,393103448	5,37784E-07
27	0,000330197	Puas	0,606896552	0,000200395	0,393103448	0,000509777
28	6,41173E-07	Puas	0,606896552	3,89126E-07	0,393103448	9,89881E-07
29	0,003764578	Puas	0,606896552	0,00228471	0,393103448	0,005811981
30	4,6865E-06	Sangat Puas	0,365517241	1,713E-06	0,634482759	2,69983E-06
31	2,38519E-13	Sangat Puas	0,365517241	8,71828E-14	0,634482759	1,37408E-13
32	5,48666E-06	Puas	0,606896552	3,32983E-06	0,393103448	8,47063E-06
33	0,007981308	Puas	0,606896552	0,004843828	0,393103448	0,01232202
34	1,68525E-06	Sangat Puas	0,365517241	6,15988E-07	0,634482759	9,7085E-07
35	7,21237E-07	Puas	0,606896552	4,37717E-07	0,393103448	1,11349E-06
36	8,17475E-06	Puas	0,606896552	4,96123E-06	0,393103448	1,26207E-05
37	1,11931E-05	Puas	0,606896552	6,79302E-06	0,393103448	1,72805E-05
38	0,001466296	Puas	0,606896552	0,00088989	0,393103448	0,002263756
39	0,008384996	Puas	0,606896552	0,005088825	0,393103448	0,012945256
40	5,02437E-05	Puas	0,606896552	3,04927E-05	0,393103448	7,75693E-05
41	3,22995E-05	Puas	0,606896552	1,96025E-05	0,393103448	4,9866E-05

42	0,000505057	Puas	0,606896552	0,000306517	0,393103448	0,000779737
43	1,10939E-08	Puas	0,606896552	6,73283E-09	0,393103448	1,71274E-08
44	0,000525249	Puas	0,606896552	0,000318772	0,393103448	0,000810911
45	0,000227655	Puas	0,606896552	0,000138163	0,393103448	0,000351467
46	2,46226E-06	Sangat Puas	0,365517241	8,99998E-07	0,634482759	1,41848E-06
47	2,46226E-06	Sangat Puas	0,365517241	8,99998E-07	0,634482759	1,41848E-06
48	2,46226E-06	Sangat Puas	0,365517241	8,99998E-07	0,634482759	1,41848E-06
49	1,64151E-06	Sangat Puas	0,365517241	5,99999E-07	0,634482759	9,4565E-07
50	2,22398E-06	Sangat Puas	0,365517241	8,12902E-07	0,634482759	1,2812E-06
51	6,74087E-08	Puas	0,606896552	4,09101E-08	0,393103448	1,0407E-07
52	2,49014E-11	Sangat Puas	0,365517241	9,10189E-12	0,634482759	1,43454E-11
53	8,40313E-07	Puas	0,606896552	5,09983E-07	0,393103448	1,29733E-06
54	2,69629E-13	Sangat Puas	0,365517241	9,85539E-14	0,634482759	1,5533E-13
55	0,001653769	Puas	0,606896552	0,001003667	0,393103448	0,002553187
56	1,23691E-13	Sangat Puas	0,365517241	4,52113E-14	0,634482759	7,1257E-14
57	0,000392415	Puas	0,606896552	0,000238156	0,393103448	0,000605834
58	3,53637E-07	Puas	0,606896552	2,14621E-07	0,393103448	5,45967E-07
59	0,000282957	Puas	0,606896552	0,000171725	0,393103448	0,000436845
60	1,67959E-05	Puas	0,606896552	1,01934E-05	0,393103448	2,59305E-05
61	1,40546E-05	Puas	0,606896552	8,52967E-06	0,393103448	2,16983E-05
62	0,009327182	Puas	0,606896552	0,005660635	0,393103448	0,01439986
63	0,007158481	Puas	0,606896552	0,004344457	0,393103448	0,011051689
64	1,60385E-06	Puas	0,606896552	9,73373E-07	0,393103448	2,47612E-06
65	0,009513726	Puas	0,606896552	0,005773847	0,393103448	0,014687858
66	2,30755E-05	Puas	0,606896552	1,40044E-05	0,393103448	3,56253E-05
67	1,80915E-27	Netral	0,020689655	3,74308E-29	0,979310345	3,82216E-29
68	2,74891E-11	Sangat Puas	0,365517241	1,00478E-11	0,634482759	1,58361E-11
69	3,24129E-05	Puas	0,606896552	1,96713E-05	0,393103448	5,00409E-05
70	1,29168E-06	Sangat Puas	0,365517241	4,7213E-07	0,634482759	7,44118E-07

Tabel 5.74 Hasil Perhitungan Probabilitas Posterior Penerbangan Surabaya – Jakarta (Lanjutan)

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
71	5,0568E-06	Puas	0,606896552	3,06895E-06	0,393103448	7,80699E-06
72	2,15641E-14	Sangat Puas	0,365517241	7,88204E-15	0,634482759	1,24228E-14
73	0,000371682	Puas	0,606896552	0,000225573	0,393103448	0,000573826
74	8,14196E-05	Puas	0,606896552	4,94133E-05	0,393103448	0,0001257
75	4,15109E-06	Puas	0,606896552	2,51928E-06	0,393103448	6,4087E-06
76	0,000836347	Puas	0,606896552	0,000507576	0,393103448	0,001291203
77	4,21942E-05	Puas	0,606896552	2,56075E-05	0,393103448	6,51419E-05
78	0,002283186	Puas	0,606896552	0,001385658	0,393103448	0,003524919
79	1,64151E-06	Sangat Puas	0,365517241	5,99999E-07	0,634482759	9,4565E-07
80	1,64151E-06	Sangat Puas	0,365517241	5,99999E-07	0,634482759	9,4565E-07
81	1,50672E-06	Puas	0,606896552	9,14426E-07	0,393103448	2,32617E-06
82	1,46379E-05	Puas	0,606896552	8,88371E-06	0,393103448	2,25989E-05
83	0,006770036	Puas	0,606896552	0,004108711	0,393103448	0,010451985
84	0,000855642	Puas	0,606896552	0,000519286	0,393103448	0,001320991
85	0,000223652	Puas	0,606896552	0,000135734	0,393103448	0,000345287
86	2,4666E-12	Sangat Puas	0,365517241	9,01586E-13	0,634482759	1,42098E-12
87	8,7852E-07	Sangat Puas	0,365517241	3,21114E-07	0,634482759	5,06104E-07
88	2,66025E-09	Sangat Puas	0,365517241	9,72366E-10	0,634482759	1,53253E-09
89	3,33619E-06	Puas	0,606896552	2,02472E-06	0,393103448	5,15061E-06
90	1,87519E-06	Puas	0,606896552	1,13804E-06	0,393103448	2,89502E-06
91	0,002007415	Puas	0,606896552	0,001218293	0,393103448	0,003099166
92	4,95821E-07	Puas	0,606896552	3,00912E-07	0,393103448	7,65478E-07

93	0,000129097	Puas	0,606896552	7,83485E-05	0,393103448	0,000199308
94	0,000198123	Puas	0,606896552	0,00012024	0,393103448	0,000305874
95	8,51315E-07	Puas	0,606896552	5,1666E-07	0,393103448	1,31431E-06
96	2,40721E-11	Sangat Puas	0,365517241	8,79876E-12	0,634482759	1,38676E-11
97	0,001939439	Puas	0,606896552	0,001177039	0,393103448	0,002994222
98	0,006418997	Puas	0,606896552	0,003895667	0,393103448	0,00991003
99	7,05137E-07	Puas	0,606896552	4,27945E-07	0,393103448	1,08863E-06
100	4,41258E-13	Sangat Puas	0,365517241	1,61287E-13	0,634482759	2,54203E-13
101	1,73943E-06	Puas	0,606896552	1,05565E-06	0,393103448	2,68544E-06
102	1,6192E-06	Sangat Puas	0,365517241	5,91844E-07	0,634482759	9,32798E-07
103	3,00465E-07	Puas	0,606896552	1,82351E-07	0,393103448	4,63875E-07
104	7,24038E-10	Sangat Puas	0,365517241	2,64648E-10	0,634482759	4,17109E-10
105	9,49333E-08	Puas	0,606896552	5,76147E-08	0,393103448	1,46564E-07
106	1,52899E-14	Sangat Puas	0,365517241	5,58872E-15	0,634482759	8,80831E-15
107	9,50404E-07	Puas	0,606896552	5,76797E-07	0,393103448	1,46729E-06
108	2,05773E-06	Sangat Puas	0,365517241	7,52135E-07	0,634482759	1,18543E-06
109	8,18467E-07	Puas	0,606896552	4,96725E-07	0,393103448	1,2636E-06
110	8,83979E-05	Puas	0,606896552	5,36484E-05	0,393103448	0,000136474
111	2,05773E-06	Sangat Puas	0,365517241	7,52135E-07	0,634482759	1,18543E-06
112	2,30105E-11	Sangat Puas	0,365517241	8,41075E-12	0,634482759	1,32561E-11
113	6,52691E-12	Sangat Puas	0,365517241	2,3857E-12	0,634482759	3,76007E-12
114	1,07986E-06	Puas	0,606896552	6,55363E-07	0,393103448	1,66715E-06
115	6,05244E-07	Puas	0,606896552	3,6732E-07	0,393103448	9,34412E-07
116	3,57998E-11	Sangat Puas	0,365517241	1,30854E-11	0,634482759	2,06238E-11
117	1,87111E-05	Puas	0,606896552	1,13557E-05	0,393103448	2,88873E-05
118	0,000103214	Puas	0,606896552	6,26401E-05	0,393103448	0,000159348
119	4,59905E-07	Puas	0,606896552	2,79115E-07	0,393103448	7,10028E-07
120	0,000434955	Puas	0,606896552	0,000263973	0,393103448	0,000671509

Tabel 5.74 Hasil Perhitungan Probabilitas Posterior Penerbangan Surabaya – Jakarta (Lanjutan)

Koresponden	P (At|K)	Klasifikasi	Probabilitas Prior P(K)	P(K) P(At|K)	P (T)	Probabilitas Posterior
121	2,88616E-06	Sangat Puas	0,365517241	1,05494E-06	0,634482759	1,66268E-06
122	1,97062E-06	Puas	0,606896552	1,19596E-06	0,393103448	3,04237E-06
123	1,11726E-08	Sangat Puas	0,365517241	4,0838E-09	0,634482759	6,43642E-09
124	2,85119E-09	Sangat Puas	0,365517241	1,04216E-09	0,634482759	1,64253E-09
125	3,70409E-06	Puas	0,606896552	2,248E-06	0,393103448	5,7186E-06
126	4,58363E-19	Tidak Puas	0,006896552	3,16113E-21	0,993103448	3,18308E-21
127	1,82581E-06	Sangat Puas	0,365517241	6,67366E-07	0,634482759	1,05183E-06
128	8,8898E-06	Sangat Puas	0,365517241	3,24937E-06	0,634482759	5,1213E-06
129	8,8898E-06	Sangat Puas	0,365517241	3,24937E-06	0,634482759	5,1213E-06
130	1,12483E-05	Sangat Puas	0,365517241	4,11145E-06	0,634482759	6,48001E-06
131	0,000908867	Puas	0,606896552	0,000551588	0,393103448	0,001403163
132	5,11798E-06	Sangat Puas	0,365517241	1,87071E-06	0,634482759	2,9484E-06
133	1,20382E-07	Puas	0,606896552	7,30592E-08	0,393103448	1,85852E-07
134	5,08552E-07	Sangat Puas	0,365517241	1,85885E-07	0,634482759	2,9297E-07
135	2,58158E-06	Sangat Puas	0,365517241	9,43612E-07	0,634482759	1,48721E-06
136	6,3924E-14	Sangat Puas	0,365517241	2,33653E-14	0,634482759	3,68258E-14
137	1,62274E-06	Puas	0,606896552	9,84836E-07	0,393103448	2,50528E-06
138	0,000369627	Puas	0,606896552	0,000224325	0,393103448	0,000570652
139	5,07048E-06	Sangat Puas	0,365517241	1,85335E-06	0,634482759	2,92103E-06
140	2,94491E-06	Sangat Puas	0,365517241	1,07642E-06	0,634482759	1,69653E-06
141	2,32455E-06	Puas	0,606896552	1,41076E-06	0,393103448	3,58878E-06
142	7,67393E-11	Sangat Puas	0,365517241	2,80495E-11	0,634482759	4,42085E-11
143	1,40648E-05	Puas	0,606896552	8,53589E-06	0,393103448	2,17141E-05
144	6,3665E-08	Sangat Puas	0,365517241	2,32706E-08	0,634482759	3,66766E-08
145	2,80444E-08	Puas	0,606896552	1,70201E-08	0,393103448	4,32966E-08

Berdasarkan hasil perhitungan diatas dapat dibuat rangkuman yang dapat dilihat pada tabel 5.75.

Tabel 5.75 Rangkuman Probabilitas Posterior Penerbangan Surabaya – Jakarta

Klasifikasi	Probabilitas Prior	Probabilitas Posterior
Sangat Puas	0,365517241	4,6026E-05
Puas	0,606896552	0,13830611
Netral	0,020689655	3,8249E-29
Tidak Puas	0,006896552	3,1831E-21
Sama Sekali Tidak Puas	0	0

Probabilitas prior adalah probabilitas yang didapatkan dari data awal penelitian, sedangkan probabilitas posterior adalah Probabilitas yang telah diperbaiki setelah mendapat informasi tambahan atau setelah melakukan penelitian. Penarikan kesimpulan pada metode bayes menggunakan hasil data probabilitas posterior dengan cara melihat hasil dari probabilitas posterior. Kesimpulan diambil pada saat nilai probabilitas posterior tertinggi pada klasifikasi atau dapat di rumuskan sebagai berikut :

$h_{MAP} = \arg\max P(x \mid h)\ P(h)$ .....(5,8)

Pada penelitian ini penerbangan Surabaya – Jakarta berdasarkan data tabel 5.97 dapat ditarik kesimpulan bahwa tingkat kepuasan pelayanan yang diberikan perusahaan kepada responden berada pada tingkat kepuasan puas. Hasil dari ini memiliki kesamaan dengan hasil *data mining* dengan menggunakan *software* weka sehingga kesimpulan dapat diterima.

V.5.2 Perhitungan Weka Data *Mining*

1. Penerbangan Jakarta – Surabaya

Hasil dari perhitungan data *mining* menggunakan *software* weka pada penerbangan Jakarta – Surabaya menghasilkan statistik seperti gambar 5.5, pada gambar tersebut data yang digunakan memiliki tingkat keberhasilan data dalam mengklasifikasi hasil oleh weka berada pada 95%. Dengan nilai keberhasilan yang tinggi tersebut memiliki mean absolute error yang rendah yaitu berada pada nilai 0.0291, sehingga dapat dikatakan bahwa hasil tersebut memiliki margin error yang sangat kecil. Pada hasil statistik *coverage*

of case dapat dilihat bahwa hasil berada di atas level *coverage of case* yang sudah ditentukan. Dengan melihat hasil statistik data *mining* menggunakan *software* weka dapat ditarik kesimpulan bahwa hasil data *mining* dapat diterima.

```
=== Summary ===

Correctly Classified Instances         138               95.1724 %
Incorrectly Classified Instances         7                4.8276 %
Kappa statistic                          0.9008
Mean absolute error                      0.0291
Root mean squared error                  0.1453
Relative absolute error                 11.5063 %
Root relative squared error             41.1475 %
Coverage of cases (0.95 level)          97.931  %
Mean rel. region size (0.95 level)      27.069  %
Total Number of Instances              145
```

Gambar 5.20 Statistik Data *Mining* Penerbangan Jakarta – Surabaya

Hasil klasifikasi berdasarkan data mining menggunakan *software* weka pada penerbangan Jakarta – Surabaya menghasilkan klasifikasi yang lebih spesifik jika dibandingkan dengan hasil perhitungan manual menggunakan metode bayes. Pada hasil tersebut terlihat jumlah dari tingkat kepuasan berdasarkan hubungan tingkat kepuasan yang ada. Namun hasil klasifikasi dari data mining ini sama dengan hasil dari perhitungan manual. Secara umum bahwa tingkat kepuasan pelayanan yang diberikan selama penerbangan berada pada posisi katagori puas dengan jumlah koresponden sebanyak 88 orang yang berada pada kategori tersebut, sedangkan pada kategori sangat puas sebanyak 53 koresponden. Hasil klasifikasi menggunakan *software* weka dapat dilihat pada gambar 5.21.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 48  5  0  0 |  a = Sangat Puas
  1 87  0  0 |  b = Puas
  0  1  2  0 |  c = Netral
  0  0  0  1 |  d = Tidak Puas
```

Gambar 5.21 Matriks tingkat kepuasan Penerbangan Jakarta – Surabaya

2. Penerbangan Surabaya – Jakarta

Hasil dari perhitungan data *mining* menggunakan *software* weka pada penerbangan Surabaya – Jakarta menghasilkan statistik seperti gambar 5.22, pada gambar tersebut data yang digunakan memiliki tingkat keberhasilan data dalam mengklasifikasi hasil oleh weka berada pada 95%. Dengan nilai keberhasilan yang tinggi tersebut memiliki mean absolute error yang rendah yaitu berada pada nilai 0.0291, sehingga dapat dikatakan bahwa hasil tersebut memiliki margin error yang sangat kecil. Pada hasil statistik *coverage of case* dapat dilihat bahwa hasil berada di atas level *coverage of case* yang sudah ditentukan. Dengan melihat hasil statistik data *mining* menggunakan *software* weka dapat ditarik kesimpulan bahwa hasil data *mining* dapat diterima.

```
=== Summary ===

Correctly Classified Instances         138               95.1724 %
Incorrectly Classified Instances         7                4.8276 %
Kappa statistic                          0.9008
Mean absolute error                      0.0291
Root mean squared error                  0.1453
Relative absolute error                 11.5063 %
Root relative squared error             41.1475 %
Coverage of cases (0.95 level)          97.931  %
Mean rel. region size (0.95 level)      27.069  %
Total Number of Instances              145
```

Gambar 5.22 Statistik Data *Mining* Penerbangan Surabaya – Jakarta

Hasil klasifikasi berdasarkan data mining menggunakan *software* weka pada penerbangan Surabaya – Jakarta menghasilkan klasifikasi yang lebih spesifik jika dibandingkan dengan hasil perhitungan manual menggunakan metode bayes. Pada hasil tersebut terlihat jumlah dari tingkat kepuasan berdasarkan hubungan tingkat kepuasan yang ada. Namun hasil klasifikasi dari data mining ini sama dengan hasil dari perhitungan manual. Secara umum bahwa tingkat kepuasan pelayanan yang diberikan selama penerbangan berada pada posisi katagori puas dengan jumlah koresponden sebanyak 88 orang yang berada pada kategori tersebut, sedangkan pada kategori sangat puas sebanyak 53 koresponden. Hasil klasifikasi menggunakan *software* weka dapat dilihat pada gambar 5.23.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 48  5  0  0 |  a = Sangat Puas
  1 87  0  0 |  b = Puas
  0  1  2  0 |  c = Netral
  0  0  0  1 |  d = Tidak Puas
```

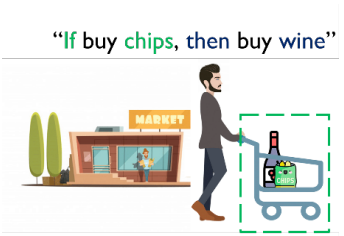
Gambar 5.23 Matriks tingkat kepuasan Penerbangan Surabaya – Jakarta

5.5 Latihan Soal

1. Tabel Transaksi

NOTRANS	ITEM_ID	QTY	NAMA_KATEGORI
1	200731154	1	AGRICULTURE
1	200034109	1	CHILDREN'S BOOKS
1	203671333	1	DIET & HEALTH
2	203900833	1	BUSINESS
2	204087933	1	CHILDREN'S BOOKS
2	203177199	1	ENTERTAINMENT
3	203538801	1	AGRICULTURE
3	203782134	1	BUSINESS
3	204049950	1	CHILDREN'S BOOKS

   a. Berapakah maksimum *Frequensi Itemsets* pada transaksi tabel transaksi di atas?
   b. Berapakah itemsets yang terbentuk jika item maksimum yang digunakan 3 item?
   c. Berapakah nilai *Support* jika seorang pelanggan membeli buku *Agriculture* dan diikuti juga membeli buku *Children's Book* pada Tabel Transaksi diatas?
   d. Berapakah nilai *Support* dari pembelian buku *Business, Agriculture* dan *Entertainment* yang dibeli secara bersamaan?

2. 

Minimum Support = 30%

Transaction ID	Itemsets
T1	
T2	
T3	
T4	
T5	
T6	

Pemilik supermarket ingin meningkatkan penjualan produk pada tahun 2022 dengan mengevaluasi penjualan produk di tahun 2021. Pada kasus Groceries data set, temukan!

a. 5 rule dengan lift tertinggi, apabila min support 0.01, confidence = 0.3 dan maxlen=4. Berikan penjelasan untuk hasil yang didapatkan!

b. Pemilik melihat penjualan whole milk tertinggi selama 2021, sehingga pemilik ingin mengetahui produk-produk apa yang dipengaruhi atau menjadi consequent. Simpulkan hasil berdasarkan 5 rule dengan lift tertinggi dan berikan penjelasan.

c. Kerjakan perhitungan Asosisi dengan menggunakan R.

3. Berdasarkan data pada jurnal Ramadani Saputra, Alexander J.P. Sibarani. Implementasi Data Mining Menggunakan Algoritma Apriori Untuk Meningkatkan Pola Penjualan Obat, Jurnal Teknik Informatika dan Sistem Informasi ISSN 2407-4322 Vol. 7, No. 2, Agustus 2020, Hal. 262-276. Buatlah Analisa Asosiasi menggunakan rumus, software Weka dan Knime dan jelaskan maksud dari jurnal tersebut.

4. Golden rule (threshold) yang digunakan adalah : 60% atau barang yang dibeli paling sedikit 3.

   Untuk mempermudah, nama-nama item di Tabel 1, disingkat dengan diambil huruf awalnya saja

   sebagai contoh :

ID Transaksi	Barang yang Dibeli
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

M = Mango

O = Onion

Dan sebagainya.

ID Transaksi	Barang yang Dibeli
T1	{M, O, N, K, E, Y }
T2	{D, O, N, K, E, Y }
T3	{M, A, K, E}
T4	{M, U, C, K, Y }
T5	{C, O, O, K, I, E}

Dalam langkah ini, misalkan ada tiga pasangan item ABC, ABD, ACD, ACE, BCD dan akan dibuatkan pasangan 4 item, carilah 2 huruf awal yang sama. Contoh :

- **AB**C dan **AB**D, menjadi ABCD
- **AC**D dan **AC**E, menjadi ACDE

Dan seterusnya.

a. Carilah tiga item yang sering dibeli bersamaan.

b. Cari asosiasi pada semua himpunan bagian yang telah dibuat dan tingkat keyakinannya.

5. Dengan menggunakan data Pistachio di atas lakukan analisis asosiasi dengan algoritma FP-Growth dengan ketentuan support=0.05, confidence=0.3, maxLength=5. Petunjuk dapat merujuk bahan materi kuliah sebelumnya untuk R Studio, boleh menggunakan software lain.

   a. Lakukan sorting rules yang terbentuk berdasarkan lift terbesar dan tunjukkan hasilnya!

   b. Apakah hasil asosiasi yang dilakukan hasilnya sejalan dengan klasifikasi yang dilakukan? Jelaskan!

BAB VI
DASAR DAN ANALISA CLUSTERING

6.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk mengenalkan definisi prinsip clustering dan aplikasinya.

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai prinsip clustering. Capaian pembelajaran lulusan adalah mahasiswa mampu menggunakan aplikasi software untuk memecahkan masalah clustering.

6.2 Konsep Clustering

Tidak seperti klasifikasi dan regresi, yang menganalisis kumpulan data berlabel kelas (pelatihan), clustering menganalisis objek data tanpa berkonsultasi dengan label kelas. Clustering dapat digunakan untuk menghasilkan label kelas untuk sekelompok data. Objek- objek tersebut dikelompokkan atau dikelompokkan berdasarkan prinsip memaksimalkan kesamaan intraclass dan meminimalkan kesamaan antar kelas. Artinya, cluster-cluster objek dibentuk agar objek-objek di dalam cluster memiliki kemiripan yang tinggi dibandingkan satu sama lain, tetapi agak berbeda dengan objek-objek dalam cluster lain. Setiap cluster yang terbentuk dapat dilihat sebagai kelas objek, dari mana aturan dapat diturunkan. Pengelompokan juga dapat memfasilitasi pembentukan taksonomi, yaitu, pengorganisasian pengamatan ke dalam hierarki kelas yang mengelompokkan peristiwa serupa bersama-sama.

Analisis klaster atau clustering adalah proses mempartisi sekumpulan objek data (atau pengamatan) ke dalam himpunan bagian. Setiap subset adalah cluster, sehingga objek dalam cluster mirip satu sama lain, namun berbeda dengan objek di cluster lain.

Kumpulan cluster yang dihasilkan dari analisis cluster dapat disebut sebagai clustering. Dalam konteks ini, metode pengelompokan yang berbeda dapat menghasilkan pengelompokan yang berbeda pada kumpulan data yang sama. Partisi tidak dilakukan oleh manusia, tetapi oleh algoritma clustering. Oleh karena itu, pengelompokan berguna karena dapat mengarah pada penemuan kelompok yang sebelumnya tidak diketahui dalam data.

Analisis cluster telah banyak digunakan dalam banyak aplikasi seperti kecerdasan bisnis, pengenalan pola gambar, pencarian Web, biologi, dan keamanan. Dalam intelijen bisnis, pengelompokan dapat digunakan untuk mengatur sejumlah besar pelanggan ke dalam kelompok, di mana pelanggan dalam suatu kelompok memiliki karakteristik serupa yang kuat. Ini memfasilitasi pengembangan strategi bisnis untuk meningkatkan manajemen hubungan pelanggan. Selain itu, pertimbangkan perusahaan konsultan dengan sejumlah besar proyek. Untuk meningkatkan manajemen proyek, pengelompokan dapat diterapkan untuk mempartisi proyek ke dalam kategori berdasarkan kesamaan sehingga audit dan diagnosis proyek (untuk meningkatkan pengiriman dan hasil proyek) dapat dilakukan secara efektif.

Dalam pengenalan citra, pengelompokan dapat digunakan untuk menemukan cluster atau "subclass" dalam sistem pengenalan karakter tulisan tangan. Misalkan kita memiliki kumpulan data digit tulisan tangan, di mana setiap digit diberi label sebagai 1, 2, 3, dan seterusnya. Perhatikan bahwa mungkin ada perbedaan besar dalam cara orang menulis angka yang sama.

Ambil nomor 2, misalnya. Beberapa orang mungkin menulisnya dengan lingkaran kecil di bagian kiri bawah, sementara yang lain mungkin tidak. Kita dapat menggunakan pengelompokan untuk menentukan subkelas untuk "2", yang masing-masing mewakili variasi cara penulisan 2. Menggunakan beberapa model berdasarkan subclass dapat meningkatkan akurasi pengenalan secara keseluruhan.

Clustering juga telah menemukan banyak aplikasi dalam pencarian Web. Misalnya, pencarian kata kunci mungkin sering menghasilkan jumlah klik yang sangat besar (yaitu, halaman yang relevan dengan pencarian) karena jumlah halaman web yang sangat banyak. Clustering dapat digunakan untuk mengatur hasil pencarian ke dalam kelompok dan menyajikan hasil dengan cara yang ringkas dan mudah diakses. Selain itu, teknik pengelompokan

telah dikembangkan untuk mengelompokkan dokumen ke dalam topik, yang biasanya digunakan dalam praktik pencarian informasi.

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat. Di mana dalam *cluster* terdiri dari kumpulan benda- benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada *cluster* lainnya. Algoritma *clustering* terdiri dari dua bagian yaitu secara hierarkis dan secara partisional. Algoritma hierarkis menemukan *cluster* secara berurutan di mana *cluster* ditetapkan sebelumnya, sedangkan alogritma partisional menentukan semua kelompok pada waktu tertentu. *Clustering* juga bisa dikatakan suatu proses di mana mengelompokan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokan pada *cluster* yang sama dan memisahkan diri dengan membentuk pola yang berbeda ke *cluster* yang berbeda. Pada ulasan mengenai *clustering*-nya, Madhulatha mengidentifikasi teknik-teknik *clustering* yang umum digunakan dan pengertiannya, beserta algoritma *clustering* yang ada, manfaat dan pengaplikasiannya, dan juga kekurangannya (Madhulatha,2012). Bataineh *et.al.* membandingkan beberapa metode *clustering* yang ada pada ulasan mereka, seperti *Fuzzy C-Mean* (FCM) dan *Subtractive Clustering,* berdasarkan validitas hasil *cluster* mereka. (Bataineh,*et.al*., 2011)

Clustering dapat memainkan peran penting dalam kehidupan sehari-hari, karena tidak bisa lepas dengan sejumlah data yang menghasilkan informasi untuk memenuhi kebutuhan hidup. Salah satu sarana yang paling penting dalam hubungan dengan data adalah untuk mengklasifikasikan atau mengelompokkan data tersebut ke dalam seperangkat kategori atau *cluster*. *Clustering* dapat ditemukan dibeberapa aplikasi yang ada di berbagai bidang. Sebagai contoh pengelompokan data yang digunakan untuk menganalisa data statistik seperti pengelompokan untuk pembelajaran mesin, *data mining*, pengenalan pola, analisis citra dan bioinformatika. Varghese *et.al.* menggunakan *clustering* sebagai metodologi untuk melakukan analisis terhadap data siswa untuk mengkarakterisasi pola performansi siswa (Varghese et.al. 2011).

Algoritma *K-means* merupakan salah satu algoritma dengan partitional, karena *K- Means* didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai *centroid* awalnya.

Algoritma *K-means* menggunakan proses secara berulang-ulang untuk mendapatkan basis data *cluster*. Dibutuhkan jumlah *cluster* awal yang diinginkan sebagai masukan dan menghasilkan titik *centroid* akhir sebagai output. Metode *K-means* akan memilih pola k sebagai titik awal *centroid* secara acak atau random. Jumlah iterasi untuk mencapai *cluster centroid* akan dipengaruhi oleh calon *cluster centroid* awal secara random. Sehingga didapat cara dalam pengembangan algoritma dengan menentukan *centroid cluster* yang dilihat dari kepadatan data awal yang tinggi agar mendapatkan kinerja yang lebih tinggi. Hung et. al. menggunakan algoritma *K-Means* ini dalam penelitian mereka untuk meningkatkan kecepatan metode *clustering.*

K-means *clustering* merupakan salah satu metode data *clustering* non-hirarki yang mengelompokan data dalam bentuk satu atau lebih *cluster*/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan *cluster*/ kelompok yang lain sehingga data yang berada dalam satu *cluster*/ kelompok memiliki tingkat variasi yang kecil (Agusta, 2007).

6.3 Studi Kasus Clustering

1. Studi Kasus *K-Means Clustering* Data COVID-19 (Indraputra R. A., Fitriana R., 2020)

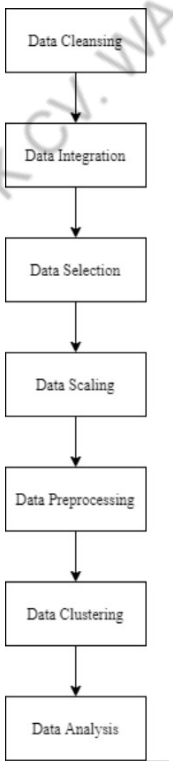
Data-data dapat ditemukan dan diakses pada *website database* seperti Kaggle, namun karena ada banyaknya data yang tersedia, maka *Big Data* seperti ini sangatlah sulit untuk dimanfaatkan secara baik. Maka, agar dapat diperoleh informasi yang berguna dan mudah untuk digunakan, dapat digunakan teknik *Data Mining* untuk mengolah data dengan jumlah sebesar ini. Salah satu teknik yang dapat digunakan adalah teknik *Clustering*, terutamanya *K-Means Clustering*, dimana data dikategorikan berdasarkan *mean* terdekat, sehingga daerah-daerah dapat dikategorikan menjadi kluster-kluster berdasarkan dampak COVID pada daerah tersebut, dan prioritisasi bantuan COVID dapat ditentukan dan diarahkan berdasarkan informasi kluster tersebut, sehingga penindakan terhadap pandemi COVID-19 ini dilakukan secara seefektif mungkin.

Data yang digunakan pada penelitian ini adalah data COVID-19 yang diperoleh dari website *database* yaitu Kaggle. Pada *database* ini

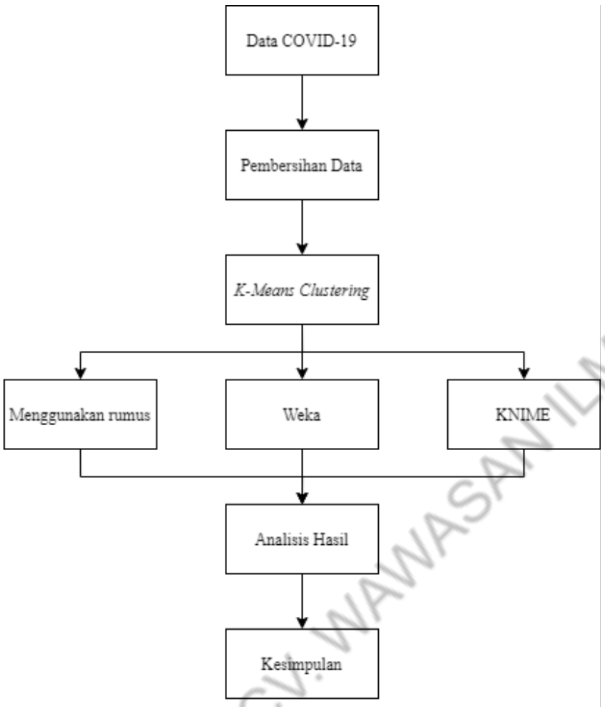
terdapat informasi yaitu nomor data, tanggal observasi, provinsi/ *state,* negara/daerah, tanggal *update* terakhir, jumlah kasus yang dikonfirmasi, jumlah kematian, dan jumlah orang yang pulih.

Metodologi pengolahan data yang digunakan adalah metodologi *data mining* yaitu *data cleasing, data integration, data selection, data scaling, data preprocessing, data clustering,* dan *data analysis*. Dalam tahap pembersihan data hal yang dilakukan adalah pembersihan data yang merupakan *outlier*, tidak lengkap, dan memilih sampel data yang akan digunakan untuk penelitian ini yaitu sebanyak 48 jumlah data.

Untuk mengolah data ini digunakan metode *K-Means Clustering* dengan tiga cara yang berbeda, yaitu menggunakan rumus dan *software Microsoft Excel*, dan menggunakan *software data mining* yaitu *Weka* dan *Knime*. Untuk penelitian ini jumlah *cluster* yang akan digunakan adalah 2 *cluster* agar hasil konsisten dan hasil perhitungan ketiga metode akan dianalisis hasilnya untuk menentukan kesimpulan dari penelitian ini.



Gambar 6.1 Flowchart Metodologi Data Mining



Gambar 6.2 Flowchart Metodologi Pengolahan Data

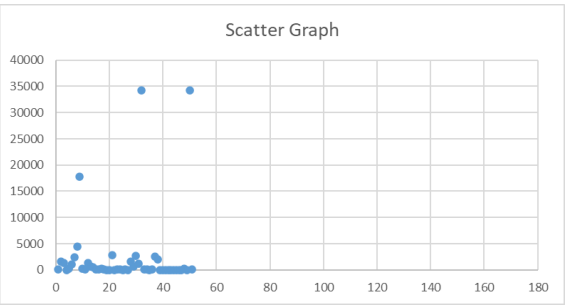
Untuk *dataset* yang diperoleh, pertama sebelum diolah perlu dilakukan pembersihan data (*data cleaning*) dan pengambilan sampel, karena ukuran data yang tersedia terlalu besar dan penelitian ini dilakukan untuk tujuan hanya untuk menguji metode *clustering* tersebut. Akan tetapi, metode ini tetap feasibel apabila seluruh *Big Data* memang mau di-*cluster*. Maka, tahap pertama dalam pengolahan data ini adalah pembersihan data yaitu menghapus subset data yang tidak menjadi perhatian dari penelitian ini, seperti tanggal observasi dan tanggal *update* terakhir. Maka, tersisa informasi yaitu nomor data, provinsi/*state*, negara/daerah, dan metode peng-*cluster*-an data ini yaitu jumlah kasus yang dikonfirmasi, jumlah kematian, dan jumlah orang yang pulih (yang berarti untuk *clustering* ini terdapat 3 variabel) Kemudian dipilih sampel data acak sebanyak 48 set data dengan variasi yang baik agar kluster terlihat dengan jelas. Diperoleh lah tabel data berupa seperti ini:

Tabel 6.1 Data Covid-19 yang telah Dibersihkan Novel Corona Virus 2019 Dataset | Kaggle. [Online] Available: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019- dataset

No	Province/ State	Country/ Region	Confirmed	Deaths	Recovered
1	Quindio	Colombia	2941	87	1905
2	Quintana Roo	Mexico	11478	1598	9485
3	Rajasthan	India	118793	1367	98812
4	Reunion	France	3501	15	2482
5	Rheinland	Germany	10246	248	9271
6	Rhode Island	US	24177	1102	0
7	Rio Grande	Brazil	67459	2355	39924
8	Rio Grande	Brazil	177485	4472	165170
9	Rio de Janeiro	Brazil	253756	17798	232036
10	Risaralda	Colombia	10239	220	7579
11	Rivne Oblast	Ukraine	11459	155	8990
12	Rondonia	Brazil	63781	1312	54687
13	Roraima	Brazil	48302	613	15324
14	Rostov Oblast	Russia	20604	460	16436
15	Ryazan Oblast	Russia	7891	49	6614
16	Saarland	Germany	3306	176	3039
17	Sachsen	Germany	6855	228	5952
18	Anhalt	Germany	2500	67	2239
19	Saga	Japan	242	0	244
20	Barthelemy	France	45	0	25
21	St. Petersburg	Russia	41359	2817	28141
22	Saint Pierre	France	12	0	6
23	Saitama	Japan	4481	100	4078
24	Republic	Russia	8807	95	6748
25	Sakhali Oblast	Russia	5347	0	3802
26	Samara Oblast	Russia	10999	177	8547
27	San Andres	Colombia	1055	10	445
28	San Luis	Mexico	22100	1601	19656
29	San Martin	Peru	16759	693	0

No	Province/ State	Country/ Region	Confirmed	Deaths	Recovered
30	Santa Catarina	Brazil	207692	2671	196530
31	Santander	Colombia	28751	1209	22261
32	Sao Paulo	Brazil	945422	34266	818593
33	Saratov Oblast	Russia	14755	95	9789
34	Sardegna	Italy	3405	145	1515
35	Saskatchewan	Canada	1830	24	1673
36	Schleswig	Germany	4498	161	4068
37	Scotland	UK	25495	2508	0
38	Sergipe	Brazil	76193	1993	70008
39	Sevastopol	Ukraine	865	25	565
40	Shaanxi	China	402	3	374
41	Shandong	China	832	7	816
42	Shanghai	China	977	7	926
43	Shanxi	China	203	0	203
44	Shiga	Japan	481	8	451
45	Shimane	Japan	138	0	137
46	Shizuoka	Japan	522	2	494
47	Sichuan	China	675	3	656
48	Sicilia	Italy	6234	303	3519

Setelah data dibersihkan, dilakukan plot data *confirmed* vs *deaths* untuk melihat apabila terdapat *outlier*, dan ternyata terdapat 3 *outlier* maka outlier tersebut perlu dibersihkan terlebih dahulu dari data. Dari *scatter graph* ini juga dapat diperkirakan berapa *cluster* yang akan terbentuk.



Gambar 6.3 Grafik scatter graph untuk mengidentifikasi terdapatnya outlier.

Sehabis *outlier* dibersihkan, untuk setiap kolom ditentukan nilai minimum, maximum dan median sebagai perkiraan kasar awal dari kluster, menggunakan fungsi =MIN(), =MAX(), dan =MEDIAN() dari *Excel*. Karena nilai maximum yaitu pada data ke-32 dibersihkan dari data karena merupakan *outlier*, maka perkiraannya hanya akan dua *cluster* yang digunakan yaitu *cluster* minimum yang berpusat pada data ke-22 (12, 0, 6) dan *cluster* median yang berpusat pada data ke-15 (7891, 49, 6614). Untuk rumus median berupa n.a. Calculating the median, 2017. [Online] Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch11/medianmediane/

$$Med\ (X) = X\left[\frac{n}{w}\right] \quad \text{.....(1)}$$

X = daftar urutan nilai pada dataset

n = jumlah data pada dataset

Kemudian, dilakukan perhitungan *Euclidean Distance* menggunakan fungsi =SUMXMY2() yang merupakan perhitungan *sum of square* pengurangan antara matriks masing-masing data dengan matriks data yang ditentukan menjadi *cluster* awal dengan rumus:

$SUMXMY2 = \sum(X - Y)^2$ .....(2)

X = nilai dari dataset 1

Y = nilai dari dataset 2

Ini digunakan untuk menentukan estimasi jarak dari data tersebut kepada *cluster*, karena metode *Clustering K-Means* ini mengkategorikan berdasarkan jarak terdekat. Data dimasukkan kedalam *cluster* 1 atau 2 berdasarkan jarak terdekatnya, menggunakan fungsi =MIN(). Dan dari hasil tersebut, diperoleh hasil *cluster* awal berupa:

Tabel 6.2 Hasil Clustering Dengan Microsoft Excel

Case	*Cluster* 1			*Cluster* 2		
	Confirmed	Deaths	Recovered	Confirmed	Deaths	Recovered
1	2941	87	1905			
2				11478	1598	9485
3				118793	1367	98812
4	3501	15	2482			
5				10246	248	9271

Case	*Cluster* 1			*Cluster* 2		
	Confirmed	Deaths	Recovered	Confirmed	Deaths	Recovered
6				24177	1102	0
7				67459	2355	39924
8				177485	4472	165170
9				253756	17798	232036
10				10239	220	7579
11				11459	155	8990
12				63781	1312	54687
13				48302	613	15324
14				20604	460	16436
15				7891	49	6614
16	3306	176	3039			
17				6855	228	5952
18	2500	67	2239			
19	242	0	244			
20	45	0	25			
21				41359	2817	28141
22	12	0	6			
23				4481	100	4078
24				8807	95	6748
25				5347	0	3802
26				10999	177	8547
27	1055	10	445			
28				22100	1601	19656
29				16759	693	0
30				207692	2671	196530
31				28751	1209	22261
33				14755	95	9789
34	3405	145	1515			
35	1830	24	1673			
36				4498	161	4068
37				25495	2508	0
38				76193	1993	70008
39	865	25	565			
40	402	3	374			

Case	*Cluster* 1			*Cluster* 2		
	Confirmed	Deaths	Recovered	Confirmed	Deaths	Recovered
41	832	7	816			
42	977	7	926			
43	203	0	203			
44	481	8	451			
45	138	0	137			
46	522	2	494			
47	675	3	656			
48				6234	303	3519
Mean	1259.5789	30.47368	957.63158	46642.67857	1657.143	37408.1071

Tabel 6.3 Kesimpulan Clustering Excel

Cluster	Jumlah Data
1	29
2	19

Untuk menghitung tingkat keakurasian *cluster*, dapat digunakan perhitungan *Sum of Square Error* (SSE). Semakin kecil nilai SSE, maka semakin akurat perkiraan *clustering* menggunakan *K-Means* ini. Rumus *Sum of Square Error* adalah (Tan,Kumar,2005)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist\ (m_i, x) \quad (3)$$

x = poin data pada *cluster* Ci

mi = poin perwakilan untuk *cluster* Ci

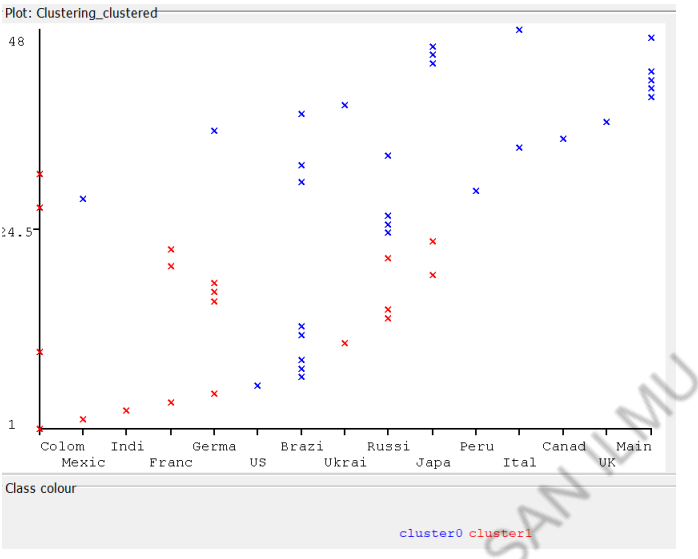
Untuk iterasi pertama ini diperoleh nilai SSE sebesar 286778721384. Maka, sebaiknya dilakukan *clustering* kembali menggunakan *Cluster K-Means* baru yang telah diperoleh yaitu (1259.5789,30.47368,957.63158) untuk *cluster* 1 dan (46642.67857,16 57.143,37408.1071)

untuk *cluster* 2, menggunakan tahap-tahap yang sama dengan tadi. Untuk iterasi kedua, diperoleh nilai SSE sebesar 183006796028. Terlihat bahwa nilai SSE mengecil secara signifikan, maka berarti iterasi *cluster* kedua ini jauh lebih akurat daripada *cluster* pertama. Untuk hasil estimasi paling optimal, sebaiknya dilakukan reiterasi hingga perubahan nilai SSE menjadi tidak signifikan lagi setelah reiterasi.

Data juga dapat diolah menggunakan teknik *clustering* menggunakan *software Weka.* Pertama, pada bagian *preprocess* keluarkan atribut yang tidak akan digunakan yaitu *date observed* dan *date last updated*. Kemudian, buka tab *cluster* dan untuk metode *clusterer* pilih metode *SimpleKMeans*. Untuk *cluster mode,* gunakan *use training set* dan klik *Start*. Maka diperoleh hasil *clustering Weka* berupa ini:

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 88.07096571388348

Initial starting points (random):

Cluster 0: 10,Risaralda,Colombia,10239,220,7579
Cluster 1: 1,Quindio,Colombia,2941,87,1905

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                       Full Data             0            1
                                   (48.0)        (29.0)       (19.0)
=============================================================================
SNo                                 24.5        30.931       14.6842
Province/State                   Quindio Rhode Island      Quindio
Country/Region                    Brazil       Brazil     Colombia
Confirmed                     47403.1042   68606.5862   15039.8947
Deaths                         1692.6042    2497.8966     463.4737
Recovered                     39254.4792   57110.7241   12000.2105
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      29 ( 60%)
1      19 ( 40%)
```

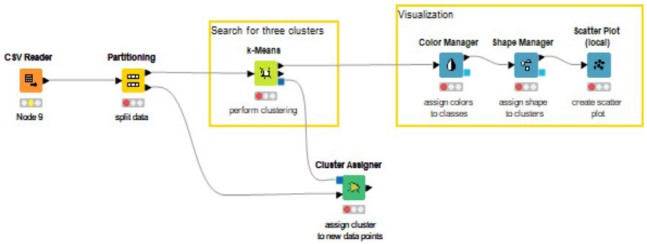
Gambar 6.4 Hasil Perhitungan Clustering Weka



Gambar 6.5 Visualisasi Clustering Weka

Untuk *clustering* menggunakan *KNIME*, pertama buka *training set* untuk *node repository Clustering*. Maka akan muncul *node CSV Reader* untuk membaca data *COVID-19* yang sudah dalam bentuk format *.csv*, yang kemudian terhubung pada *node Partitioning* untuk membagi data, yang terbuhung dengan dua *node*, yaitu *node k-Means* untuk menentukan nilai *K-Means* pada data dan melakukan *clustering* dan *node Cluster Assigner* yang menugaskan *cluster* pada data *point* baru.

Untuk visualisasi data, terdapat *node Color Manager* yang memberikan warna-warna yang berbeda pada kelas yang berbeda, *node Shape Manager* yang memberikan bentuk berbeda pada *cluster* yang berbeda, dan *node Scatter Plot* yang membuat *scatter diagram*.



Gambar 6.6 Layout node repository K-Means Clustering dengan KNIME.

Apabila seluruh *node* telah di-*setup* dengan lengkap, maka hal yang tinggal perlu dilakukan adalah memasukan data *COVID-19* pada *CSV Reader*, menentukan warna dan bentuk visualisasi yang diinginkan dan meng-*execute* seluruh *node*, dan memilih opsi *Open Views* untuk menghasilkan visualisasi *clustering* yang diperoleh.

Gambar 6.7 Visualiasi Clustering KNIME.

Kesimpulan yang didapatkan adalah sampel data ini dapat dibagi menjadi 2 *cluster*, dimana pada *cluster* 1 terdapat 29 data dan pada *cluster* 2 terdapat 19 data. Dengan informasi ini, dapat diketahui *cluster* daerah-daerah yang memerlukan penanganan COVID-19 yang lebih darurat, di mana pada sampel ini merupakan *cluster* 2 dengan jumlah terjangkit dan meninggal yang lebih tinggi dibandingkan dengan *cluster* 1.

2. Studi Kasus Clustering di Usaha Kecil Menengah Roti (Fitriana *et. al.*,2018)

Pre-processing data berfokus pada sumber data yang akan digunakan untuk keperluan pengolahan *data mining*.

a. Jenis Cacat

Jenis cacat yang dihasilkan ketika proses produksi berlangsung yaitu gosong, bantat dan hancur

Tabel 6.4 Jenis Cacat Pada Produk Roti

Jenis cacat	Pengertian
Gosong	Warna Kulit Roti Berwarna Hitam Bergaris
Bantat	Tekstur Roti Sedikit Keras
Hancur	Tekstur Roti Tidak Terbentuk

b. *Cluster* Jumlah Cacat

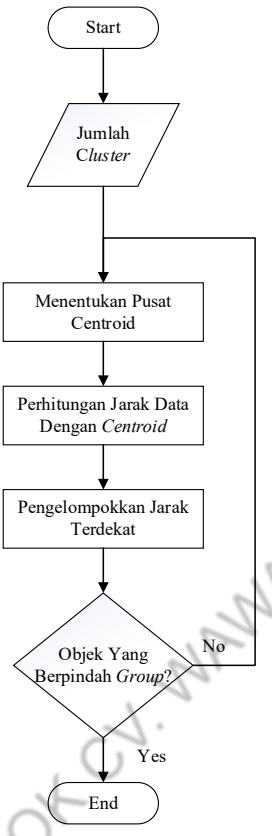
Jumlah cacat produk ditransformasi menjadi 3 klasifikasi yaitu *cluster* 1, *cluster* 2 dan *cluster* 3.

Tabel 6.5 Pengelompokkan Jumlah Cacat

Cluster	Kriteria	Range
Cluster 1	Jumlah cacat produk roti kecil	< 1000
Cluster 2	Jumlah cacat produk roti sedang	1000 < x < 2000
Cluster 3	Jumlah cacat produk roti tinggi	> 2000

Perhitungan *Clustering*

Clustering merupakan penglompokkan sejumlah data sehingga memudahkan untuk dianalisa lebih lanjut. Pengelompokkan dilakukan berdasarkan jumlah cacat yang dihasilkan. Berikut tahapan-tahapan yang harus dilakukan :



Gambar 6.8 *Flowchart Clustering K-Means*

Langkah pertama dalam perhitungan *clustering k-means* yaitu menentukan jumlah *cluster*. Dalam penelitian ini penetuan *cluster* yaitu sebanyak 3 *cluster*.

1. *Cluster* 1 : Jumlah cacat kecil yaitu untuk kerusakan < 1000 buah roti.
2. *Cluster* 2 : Jumlah cacat sedang yaitu produk roti dengan jumlah cacat 1000 – 2000 buah roti.
3. *Cluster* 3 : Jumlah cacat besar yaitu untuk kerusakan >2000 buah roti.

Pada tahap menentukan pusat *centroid* yaitu menentukan titik pusat pada data yang diperoleh. Penentuan pusat data ini dilakukan secara *random.* Perhitungan jarak data dengan *centroid* yaitu dengan menggunakan rumus sebagai berikut :

$$d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Berikut data-data yang digunakan untuk melakukan perhitungan jarak data dengan *centroid*.

Tabel 6.6 Jumlah Cacat Produk

kode Produk	Nama Produk	Gosong	Bantat	Hancur
PD-1	Coklat	2998	2380	2591
PD-2	Kelapa	1462	1249	1211
PD-3	Sarikaya	1319	1303	1244
PD-4	Mocha	1147	1082	1001
PD-5	Donat	3095	2656	2554
PD-6	Agar	657	567	518
PD-7	Sate	1382	1239	1344
PD-8	Pisang Coklat	1051	907	894
PD-9	Keju	1184	1103	1190
PD-10	Pisang Coklat Keju	2306	2074	2017
PD-11	Blueberry	117	69	35
PD-12	Keju Coklat	1925	1872	1861
PD-13	Daging	933	941	947
PD-14	Sosis	800	754	667
PD-15	Susu	1192	1072	1068
PD-16	5 Rasa	604	524	549
PD-17	Strawberry	721	649	726
PD-18	Sosis Abon	449	421	401
PD-19	KPK	566	506	512
PD-20	Set Kotak	1542	1341	1474
PD-21	Set Kecil	1200	1149	1061
PD-22	Kupas	2819	2888	2931
PD-23	Tawar	5012	4085	4109

Iterasi ke -1

1. Penentuan pusat awal *cluster*

   Untuk penentuan awal diasumsikan :

   Diambil data ke-11 sebagai pusat *cluster* ke-1 : (117,69,35)

   Diambil data ke-5 sebagai pusat *cluster* ke-2 : (3095,2656,2554)

   Diambil data ke-23 sebagai pusat *cluster* ke-3 : (5012,4085,4109)

2. Perhitungan jarak pusat *cluster*

   Untuk mengukur jarak antara data dengan pusat *cluster* yaitu dengan menggunakan rumus *Euclidean distance* sebagai berikut:

$$d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Berikut contoh perhitungan *cluster* 1 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-117)^2 + (2380-69)^2 + (2591-35)^2}$ = **4991.549621**

PD-2 = $\sqrt{(1462-117)^2 + (1249-69)^2 + (1211-35)^2}$ = **2141.121435**

Berikut contoh perhitungan *cluster* 2 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-3095)^2 + (2380-2656)^2 + (2591-2554)^2}$

= **2948796365**

PD-2 = $\sqrt{(1462-3095)^2 + (1249-2656)^2 + (1211-2554)^2}$

= **5389.230001**

Berikut contoh perhitungan *cluster* 3 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-5012)^2 + (2380-4085)^2 + (2591-4109)^2}$

= **3044.264279**

PD-2 = $\sqrt{(1462-5012)^2 + (1249-4085)^2 + (1211-4109)^2}$

= **5389.230001**

Dari hasil perhitungan diatas, dapat dirangkum seperti tabel berikut ini :

Tabel 6.7 Hasil Perhitungan *Clustering*

kode Produk	Nama Produk	Gosong	Bantat	Hancur	C1	C2	C3	Jarak Terpendek
PD-1	Coklat	2998	2380	2591	4491.549621	294.8796365	3044.264279	294.8796365
PD-2	Kelapa	1462	1249	1211	2141.121435	2539.68246	5389.230001	2141.121435
PD-3	Sarikaya	1319	1303	1244	2104.576204	2588.606768	5439.301242	2104.576204
PD-4	Mocha	1147	1082	1001	1737.879455	2946.860872	5797.921869	1737.879455
PD-5	Donat	3095	2656	2554	4680.428827	0	2852.184251	0
PD-6	Agar	657	567	518	879.143333	3801.718164	6651.137497	879.143333
PD-7	Sate	1382	1239	1344	2163.932993	2531.078426	5377.90303	2163.932993
PD-8	Pisang Coklat	1051	907	894	1520.684385	3161.097436	6010.443411	1520.684385
PD-9	Keju	1184	1103	1190	1881.932517	2815.000178	5662.726287	1881.932517
PD-10	Pisang Coklat Keju	2306	2074	2017	3569.323465	1117.861351	3967.747598	1117.861351

PD-11	Blueberry	117	69	35	0	4680.428827	7529.060831	0
PD-12	Keju Coklat	1925	1872	1861	3139.10003	1569.651235	4413.66537	1569.651235
PD-13	Daging	933	941	947	1502.658977	3193.417918	6043.278994	1502.658977
PD-14	Sosis	800	754	667	1155.481718	3527.803566	6378.390785	1155.481718
PD-15	Susu	1192	1072	1068	1796.86477	2887.673977	5737.44281	1796.86477
PD-16	5 Rasa	604	524	549	841.6590759	3843.244723	6692.143528	841.6590759
PD-17	Strawberry	721	649	726	1085.678129	3606.315155	6454.724316	1085.678129
PD-18	Sosis Abon	449	421	401	606.6992665	4078.228782	6927.851687	606.6992665
PD-19	KPK	566	506	512	787.46365	3897.191938	6746.448399	787.46365
PD-20	Set Kotak	1542	1341	1474	2391.512074	2303.786883	5149.141773	2303.786883
PD-21	Set Kecil	1200	1149	1061	1841.728807	2844.489937	5695.765445	1841.728807
PD-22	Kupas	2819	2888	2931	4861.520441	521.6598509	2762.198762	521.6598509
PD-23	Tawar	5012	4085	4109	7529.060831	2852.184251	0	0

Menentukan jarak terpendek yaitu dengan membandingkan diantara 3 *cluster* yang memiliki nilai terendah.

3. Pengelompokan Data

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat. Hasil jarak terdekat antara data dengan pusat *cluster* dapat dilihat pada tabel matrix berikut :

Tabel 6.8 Tabel Matrix *Clustering*

No	C1	C2	C3
1	0	1	0
2	1	0	0
3	1	0	0
4	1	0	0
5	0	1	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	0	1	0
11	1	0	0
12	0	1	0
13	1	0	0
14	1	0	0
15	1	0	0
16	1	0	0
17	1	0	0
18	1	0	0
19	1	0	0
20	0	1	0
21	1	0	0
22	0	1	0
23	0	0	1

Angka 1 menunjukkan bahwa data tersebut memiliki jarak terdekat dengan pusat *cluster*. Setelah mengetahui jarak terdekat dengan pusat *cluster*, langkah selanjutnya yaitu dengan melakukan perhitungan *cluster* yang baru seperti pada tabel berikut :

Tabel 6.9 Perhitungan *Cluster* Baru

kode Produk	Nama Produk	Gosong	Bantat	Hancur	Cluster Baru		
					C1	C2	C3
PD-1	Coklat	2998	2380	2591	924	2447.5	5012
PD-2	Kelapa	1462	1249	1211	845.9375	2201.833333	4085
PD-3	Sarikaya	1319	1303	1244	835.5	2238	4109
PD-4	Mocha	1147	1082	1001			
PD-5	Donat	3095	2656	2554			
PD-6	Agar	657	567	518			
PD-7	Sate	1382	1239	1344			
PD-8	Pisang Coklat	1051	907	894			
PD-9	Keju	1184	1103	1190			
PD-10	Pisang Coklat Keju	2306	2074	2017			
PD-11	Blueberry	117	69	35			
PD-12	Keju Coklat	1925	1872	1861			
PD-13	Daging	933	941	947			
PD-14	Sosis	800	754	667			
PD-15	Susu	1192	1072	1068			
PD-16	5 Rasa	604	524	549			
PD-17	Strawberry	721	649	726			
PD-18	Sosis Abon	449	421	401			
PD-19	KPK	566	506	512			
PD-20	Set Kotak	1542	1341	1474			
PD-21	Set Kecil	1200	1149	1061			
PD-22	Kupas	2819	2888	2931			
PD-23	Tawar	5012	4085	4109			

Untuk menentukan *cluster* baru dapat dilakukan dengan menggunakan perhitungan sebagai berikut :

- ***Cluster* 1 (Gosong)**

  $= \left(\frac{1462+1319+1147+657+1382+1051+1184+117+933+800+1192+604+721+449+566+1200}{16}\right)$

  = 924

- ***Cluster* 1 (Bantat)**

  $= \left(\frac{1249+1303+1082+567+1239+907+1103+69+941+754+1072+524+649+421+506+1149}{16}\right)$

  = 845.9375

- ***Cluster* 1 (Hancur)**

  $= \left(\frac{1211+1244+1001+518+1344+894+1190+35+947+667+1068+549+726+401+512+1061}{16}\right)$

  = 835.5

- ***Cluster* 2 (Gosong)** $= \left(\frac{2998+3095+2306+1925+1542+2819}{6}\right) =$ **2447.5**

- ***Cluster* 2 (Bantat)** $= \left(\frac{23880+2656+2074+1872+1341+2888}{6}\right) =$ **2201.833333**

- ***Cluster* 2 (Hancur)** $= \left(\frac{2591+2554+2017+1861+1474+2931}{6}\right) =$ **2238**

- ***Cluster* 3 (Gosong) =** $\left(\frac{5012}{1}\right)$ **= 5012**

- ***Cluster* 3 (Bantat) =** $\left(\frac{4085}{1}\right)$ **= 4085**

- ***Cluster* 3 (Hancur) =** $\left(\frac{4109}{1}\right)$ **= 4109**

1. Penentuan pusat awal *cluster* yaitu sesuai dengan perhitungan di atas, sehingga disimpulkan pusat *cluster* yang baru yaitu :

   Cluster baru yang ke-1 = (924, 825.94, 835.5)

   Cluster baru yang ke-2 = (2447.5, 2201.8, 2238)

   Cluster baru yang ke-3 = (5012, 4085, 4109)

   Perhitungan jarak terpendek antara data dengan pusat *cluster* dapat dilakukan dengan menggunakan rumus :

   Berikut contoh perhitungan *cluster* 1 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-924)^2+(2380-845.94)^2+(2591-835.5)^2}$

= **3120.353186**

PD-2 = $\sqrt{(1462-924)^2+(1249-845.94)^2+(1211-835.5)^2}$

= **770.0023564**

Berikut perhitungan contoh *cluster* 2 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-2447.5)^2+(2380-2201.8)^2+(2591-2238)^2}$

= **677.7924543**

PD-2 = $\sqrt{(1462-2447.5)^2+(1249-2201.8)^2+(1211-2238)^2}$

= **1712.842845**

Berikut perhitungan contoh *cluster* 3 untuk mengukur jarak antara data dengan pusat *cluster*.

PD-1 = $\sqrt{(2998-5012)^2+(2380-4085)^2+(2591-4109)^2}$

= **3044.264279**

$$PD\text{-}2 = \sqrt{(1462-5012)^2 + (1249-4085)^2 + (1211-4109)^2}$$

$$= \mathbf{5389.230001}$$

Dari hasil perhitungan sebelumnya, diperoleh hasil perhitungan *clustering* untuk menentukan jarak terpendek antara data dengan pusat *cluster* sebagai berikut :

Tabel 6.10 Hasil Perhitungan *Clustering* Iterasi 2

kode Produk	Nama Produk	Gosong	Bantat	Hancur	C1	C2	C3	Jarak Terpendek
PD-1	Coklat	2998	2380	2591	3120.353186	677.7924543	3044.264279	677.7924543
PD-2	Kelapa	1462	1249	1211	770.0023564	1712.842845	5389.230001	770.0023564
PD-3	Sarikaya	1319	1303	1244	729.2485028	1751.984478	5439.301242	729.2485028
PD-4	Mocha	1147	1082	1001	364.4787427	2115.536798	5797.921869	364.4787427
PD-5	Donat	3095	2656	2554	3307.991763	851.6922044	2852.184251	851.6922044
PD-6	Agar	657	567	518	499.9013692	2972.704203	6651.137497	499.9013692
PD-7	Sate	1382	1239	1344	789.1985675	1691.618833	5377.90303	789.1985675
PD-8	Pisang Coklat	1051	907	894	152.5774522	2330.910039	6010.443411	152.5774522
PD-9	Keju	1184	1103	1190	509.2655289	1975.39134	5662.726287	509.2655289
PD-10	Pisang Coklat Keju	2306	2074	2017	2194.083807	291.898289	3967.747598	291.898289
PD-11	Blueberry	117	69	35	1376.837365	3851.417567	7529.060831	1376.837365
PD-12	Keju Coklat	1925	1872	1861	1762.513973	723.8268286	4413.66537	723.8268286
PD-13	Daging	933	941	947	146.7996216	2355.863312	6043.278994	146.7996216
PD-14	Sosis	800	754	667	228.5186074	2697.872979	6378.390785	228.5186074
PD-15	Susu	1192	1072	1068	420.6952625	2054.678469	5737.44281	420.6952625
PD-16	5 Rasa	604	524	549	536.773699	3011.036025	6692.143528	536.773699
PD-17	Strawberry	721	649	726	303.2880296	2770.9633	6454.724316	303.2880296
PD-18	Sosis Abon	449	421	401	771.3540879	3246.527162	6927.851687	771.3540879
PD-19	KPK	566	506	512	590.2319492	3065.121359	6746.448399	590.2319492
PD-20	Set Kotak	1542	1341	1474	1017.198667	1464.465868	5149.141773	1017.198667
PD-21	Set Kecil	1200	1149	1061	467.8387852	2012.471932	5695.765445	467.8387852
PD-22	Kupas	2819	2888	2931	3485.995482	1043.592806	2762.198762	1043.592806
PD-23	Tawar	5012	4085	4109	6157.846387	3691.018551	0	0

Dari tabel diatas dapat disimpulkan bahwa jarak terpendek antara data dengan pusat *cluster* yaitu ditunjukkan dengan tabel matrix sebagai berikut :

Tabel 6.11 Tabel Matrix *Clustering* Iterasi 2

No	C1	C2	C3
1	0	1	0
2	1	0	0
3	1	0	0
4	1	0	0
5	0	1	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	0	1	0
11	1	0	0
12	0	1	0
13	1	0	0
14	1	0	0
15	1	0	0
16	1	0	0
17	1	0	0

18	1	0	0
19	1	0	0
20	1	0	0
21	1	0	0
22	0	1	0
23	0	0	1

Dari tabel matrix yang diperoleh, dapat dilakukan perhitungan untuk mennentukan pusat *cluster* yang baru. Penentuan *cluster* yang baru akan terus berlanjut apabila matrix pada tabel matrix terbaru dan tabel matrix terdahulu dibandingkan tidak sama. Karena pada matrix terbaru dan terdahulu tidak memiliki kemiripan, maka dilakukan pehitungan kembali untuk menentukan pusat *cluster* yang baru seperti berikut:

- ***Cluster* 1 (Gosong)**

  $= \left(\frac{1462+1319+1147+657+1382+1051+1184+117+933+800+1192+604+721+449+566+1542+1200}{16}\right)$

  = 960.3529412
- ***Cluster* 1 (Bantat)**

  $= \left(\frac{1249+1303+1082+567+1239+907+1103+69+941+754+1072+524+649+421+506+1341+1149}{16}\right)$

  = 875.0588235
- ***Cluster* 1 (Hancur)**

  $= \left(\frac{1211+1244+1001+518+1344+894+1190+35+947+667+1068+549+726+401+512+1474+1061}{16}\right)$

  = 873.0588235
- ***Cluster* 2 (Gosong)** $= \left(\frac{2998+3095+2306+1925+2819}{6}\right)$ **= 2628.6**
- ***Cluster* 2 (Bantat)** $= \left(\frac{23880+2656+2074+1872+2888}{6}\right)$ **= 2374**
- ***Cluster* 2 (Hancur)** $= \left(\frac{2591+2554+2017+1861+2931}{6}\right)$ **= 2390.8**
- ***Cluster* 3 (Gosong)** $= \left(\frac{5012}{1}\right)$ **= 5012**
- ***Cluster* 3 (Bantat)** $= \left(\frac{4085}{1}\right)$ **= 4085**
- ***Cluster* 3 (Hancur)** $= \left(\frac{4109}{1}\right)$ **= 4109**

Dari perhitungan penentuan pusat *cluster* yang telah dilakukan, maka dapat dilihat hasil perhitungan pada tabel berikut :

Tabel 6.12 Perhitungan *Cluster* Baru Iterasi 2

kode Produk	Nama Produk	Gosong	Bantat	Hancur	Cluster Baru		
					C1	C2	C3
PD-1	Coklat	2998	2380	2591	960.3529412	2628.6	5012
PD-2	Kelapa	1462	1249	1211	875.0588235	2374	4085
PD-3	Sarikaya	1319	1303	1244	873.0588235	2390.8	4109
PD-4	Mocha	1147	1082	1001			
PD-5	Donat	3095	2656	2554			
PD-6	Agar	657	567	518			
PD-7	Sate	1382	1239	1344			
PD-8	Pisang Coklat	1051	907	894			
PD-9	Keju	1184	1103	1190			
PD-10	Pisang Coklat Keju	2306	2074	2017			
PD-11	Blueberry	117	69	35			
PD-12	Keju Coklat	1925	1872	1861			
PD-13	Daging	933	941	947			
PD-14	Sosis	800	754	667			
PD-15	Susu	1192	1072	1068			
PD-16	5 Rasa	604	524	549			
PD-17	Strawberry	721	649	726			
PD-18	Sosis Abon	449	421	401			
PD-19	KPK	566	506	512			
PD-20	Set Kotak	1542	1341	1474			
PD-21	Set Kecil	1200	1149	1061			
PD-22	Kupas	2819	2888	2931			
PD-23	Tawar	5012	4085	4109			

1. Penentuan pusat awal *cluster* yang baru yaitu :

   Cluster baru yang ke-1 = (960.35, 875.06, 873.06)

   Cluster baru yang ke-2 = (2628.6, 2374, 2390.8)

   Cluster baru yang ke-3 = (5012, 4085, 4109)

   Berikut contoh hasil perhitungan penentuan pusat awal *cluster* 1 yang baru :

PD-1 = $\sqrt{(2998-960.35)^2+(2380-875.06)^2+(2591-873.06)^2}$

= **3060.747518**

PD-2 = $\sqrt{(1462-960.35)^2+(1249-875.06)^2+(1211-873.06)^2}$

= **711.116034**

Berikut contoh hasil perhitungan penentuan pusat awal *cluster* 2 yang baru :

$$PD\text{-}1 = \sqrt{(2998-2628.6)^2 + (2380-2374)^2 + (2591-2390.8)^2}$$

$$= \mathbf{420.205188}$$

$$PD\text{-}2 = \sqrt{(1462-2628.6)^2 + (1249-2374)^2 + (1211-2390.8)^2}$$

$$= \mathbf{2004.62181}$$

Berikut hasil perhitungan penentuan pusat awal *cluster* 3 yang baru :

$$PD\text{-}1 = \sqrt{(2998-5012)^2 + (2380-4085)^2 + (2591-4109)^2}$$

$$= \mathbf{3044.264279}$$

$$PD\text{-}2 = \sqrt{(1462-5012)^2 + (1249-4085)^2 + (1211-4109)^2}$$

$$= \mathbf{5389.230001}$$

Dari hasil perhitungan tersebut dapat diperoleh hasil perhitungan *clustering* seperti berikut :

Tabel 6.13 Hasil Perhitungan *Clustering* Iterasi 3

kode Produk	Nama Produk	Gosong	Bantat	Hancur	C1	C2	C3	Jarak Terpendek
PD-1	Coklat	2998	2380	2591	3060.747518	420.205188	3044.264279	420.205188
PD-2	Kelapa	1462	1249	1211	711.116034	2004.62181	5389.230001	711.116034
PD-3	Sarikaya	1319	1303	1244	670.3422407	2043.83057	5439.301242	670.3422407
PD-4	Mocha	1147	1082	1001	306.6442886	2407.477227	5797.921869	306.6442886
PD-5	Donat	3095	2656	2554	3248.697089	568.9351457	2852.184251	568.9351457
PD-6	Agar	657	567	518	559.4551044	3264.940336	6651.137497	559.4551044
PD-7	Sate	1382	1239	1344	729.4004482	1984.44622	5377.90303	729.4004482
PD-8	Pisang Coklat	1051	907	894	98.36493735	2623.227211	6010.443411	98.36493735
PD-9	Keju	1184	1103	1190	449.9187662	2268.089681	5662.726287	449.9187662
PD-10	Pisang Coklat Keju	2306	2074	2017	2134.67266	577.7518498	3967.747598	577.7518498
PD-11	Blueberry	117	69	35	1436.425286	4143.784888	7529.060831	1436.425286
PD-12	Keju Coklat	1925	1872	1861	1703.074698	1013.777589	4413.66537	1013.777589
PD-13	Daging	933	941	947	102.7799578	2648.227105	6043.278994	102.7799578
PD-14	Sosis	800	754	667	287.7994845	2989.927156	6378.390785	287.7994845
PD-15	Susu	1192	1072	1068	361.1762023	2347.088281	5737.44281	361.1762023
PD-16	5 Rasa	604	524	549	596.023353	3303.593861	6692.143528	596.023353
PD-17	Strawberry	721	649	726	360.5810862	3063.677822	6454.724316	360.5810862
PD-18	Sosis Abon	449	421	401	830.9577478	3538.950296	6927.851687	830.9577478
PD-19	KPK	566	506	512	649.6784832	3357.622998	6746.448399	649.6784832
PD-20	Set Kotak	1542	1341	1474	957.3634516	1757.359041	5149.141773	957.3634516
PD-21	Set Kecil	1200	1149	1061	409.6295482	2304.320073	5695.765445	409.6295482
PD-22	Kupas	2819	2888	2931	3426.605165	769.5870321	2762.198762	769.5870321
PD-23	Tawar	5012	4085	4109	6098.432712	3400.048206	0	0

Penentuan jarak terpendek dapat diringkas dengan menggunakan tabel matrix seperti berikut :

Tabel 6.14 Tabel Matrix *Clustering* Iterasi 3

No	C1	C2	C3
1	0	1	0
2	1	0	0
3	1	0	0
4	1	0	0
5	0	1	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	0	1	0
11	1	0	0
12	0	1	0
13	1	0	0
14	1	0	0
15	1	0	0
16	1	0	0
17	1	0	0
18	1	0	0
19	1	0	0
20	1	0	0
21	1	0	0
22	0	1	0
23	0	0	1

Tabel matrix yang telah diperoleh kemudian dibandingkan dengan tabel matrix yang terdahulu. Setelah dibandingkan, ternyata isi dari tabel matrix sudah sama dengan tabel sebelumnya, sehingga hasil akhir dari perhitungan *clustering* yaitu sebagai berikut :

Tabel 6.15 Hasil *Clustering* Akhir

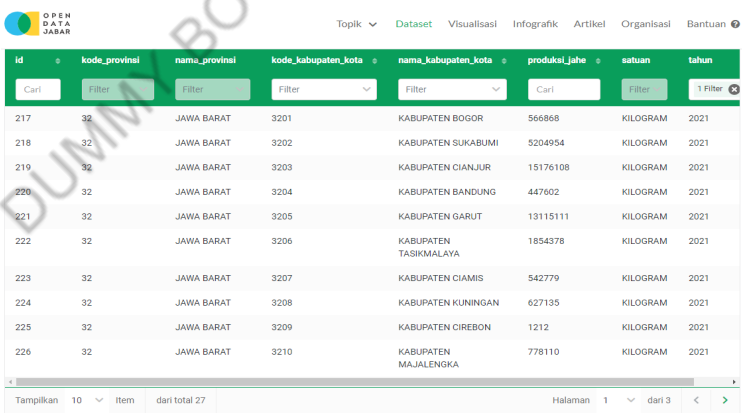
kode Produk	Nama Produk	Gosong	Bantat	Hancur	Cluster
PD-1	Coklat	2998	2380	2591	Cluster 3
PD-2	Kelapa	1462	1249	1211	Cluster 2
PD-3	Sarikaya	1319	1303	1244	Cluster 2
PD-4	Mocha	1147	1082	1001	Cluster 2
PD-5	Donat	3095	2656	2554	Cluster 3
PD-6	Agar	657	567	518	Cluster 1
PD-7	Sate	1382	1239	1344	Cluster 2
PD-8	Pisang Coklat	1051	907	894	Cluster 2
PD-9	Keju	1184	1103	1190	Cluster 2
PD-10	Pisang Coklat Keju	2306	2074	2017	Cluster 3
PD-11	Blueberry	117	69	35	Cluster 1
PD-12	Keju Coklat	1925	1872	1861	Cluster 2

PD-13	Daging	933	941	947	Cluster 1
PD-14	Sosis	800	754	667	Cluster 1
PD-15	Susu	1192	1072	1068	Cluster 2
PD-16	5 Rasa	604	524	549	Cluster 1
PD-17	Strawberry	721	649	726	Cluster 1
PD-18	Sosis Abon	449	421	401	Cluster 1
PD-19	KPK	566	506	512	Cluster 1
PD-20	Set Kotak	1542	1341	1474	Cluster 2
PD-21	Set Kecil	1200	1149	1061	Cluster 2
PD-22	Kupas	2819	2888	2931	Cluster 3
PD-23	Tawar	5012	4085	4109	Cluster 3

Dari hasil *clustering* diatas dapat dilihat bahwa yang termasuk *clustering* 3 yang merupakan cacat terbesar dengan jumlah cacat >2000 buah roti yaitu roti coklat, donat, pisang coklat keju, kupas dan roti tawar.

3. Studi Kasus Clustering Tanaman Obat (Hidayat, Fitriana,2022)

Data hasil produksi berbagai tanaman obat merupakan hasil dari proses bisnis yang telah dilakukan pada sektor pertanian pada wilayah kabupaten dan kota di provinsi Jawa Barat, yaitu tanaman serai, jeruk nipis sambiloto, lidah buaya, mahkota dewa, mengkudu, kapulaga, temu kunci, temu ireng, temulawak, lempuyang, kunyit, kencur, lengkuas dan jahe. Data hasil produksi diperoleh dari situs web https://opendata.jabarprov.go.id yang bersumber dari Dinas Tanaman Pangan dan Hortikultura Provinsi Jawa Barat.

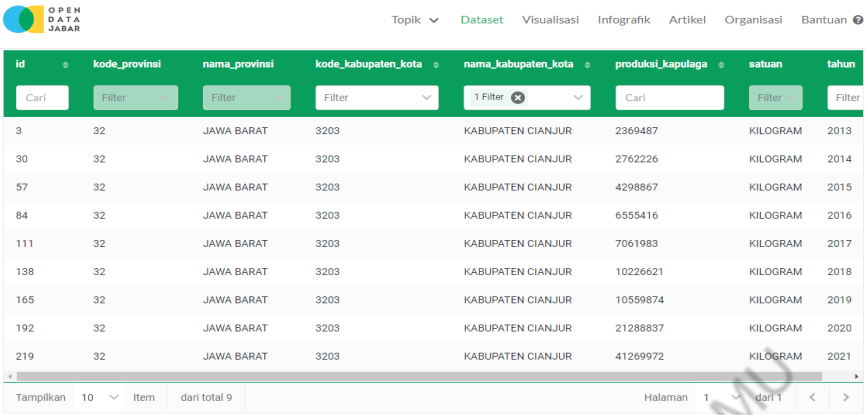

OPEN DATA JABAR

Topik Dataset Visualisasi Infografik Artikel Organisasi Bantuan

id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	produksi_jahe	satuan	tahun
Cari	Filter	Filter	Filter	Filter	Cari	Filter	1 Filter
217	32	JAWA BARAT	3201	KABUPATEN BOGOR	566868	KILOGRAM	2021
218	32	JAWA BARAT	3202	KABUPATEN SUKABUMI	5204954	KILOGRAM	2021
219	32	JAWA BARAT	3203	KABUPATEN CIANJUR	15176108	KILOGRAM	2021
220	32	JAWA BARAT	3204	KABUPATEN BANDUNG	447602	KILOGRAM	2021
221	32	JAWA BARAT	3205	KABUPATEN GARUT	13115111	KILOGRAM	2021
222	32	JAWA BARAT	3206	KABUPATEN TASIKMALAYA	1854378	KILOGRAM	2021
223	32	JAWA BARAT	3207	KABUPATEN CIAMIS	542779	KILOGRAM	2021
224	32	JAWA BARAT	3208	KABUPATEN KUNINGAN	627135	KILOGRAM	2021
225	32	JAWA BARAT	3209	KABUPATEN CIREBON	1212	KILOGRAM	2021
226	32	JAWA BARAT	3210	KABUPATEN MAJALENGKA	778110	KILOGRAM	2021

Tampilkan 10 Item dari total 27 Halaman 1 dari 3

Gambar 6.9 Sampel Data Produksi Jahe di Provinsi Jawa Barat (Dinas Tanaman Pangan dan Hortikultura Provinsi Jawa Barat, 2022)



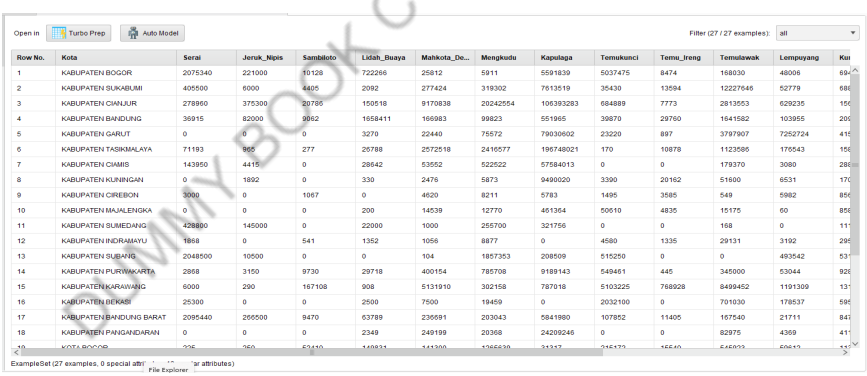
OPEN DATA JABAR — Topik | Dataset | Visualisasi | Infografik | Artikel | Organisasi | Bantuan

id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	produksi_kapulaga	satuan	tahun
Cari	Filter	Filter	Filter	1 Filter	Cari	Filter	Filter
3	32	JAWA BARAT	3203	KABUPATEN CIANJUR	2369487	KILOGRAM	2013
30	32	JAWA BARAT	3203	KABUPATEN CIANJUR	2762226	KILOGRAM	2014
57	32	JAWA BARAT	3203	KABUPATEN CIANJUR	4298867	KILOGRAM	2015
84	32	JAWA BARAT	3203	KABUPATEN CIANJUR	6555416	KILOGRAM	2016
111	32	JAWA BARAT	3203	KABUPATEN CIANJUR	7061983	KILOGRAM	2017
138	32	JAWA BARAT	3203	KABUPATEN CIANJUR	10226621	KILOGRAM	2018
165	32	JAWA BARAT	3203	KABUPATEN CIANJUR	10559874	KILOGRAM	2019
192	32	JAWA BARAT	3203	KABUPATEN CIANJUR	21288837	KILOGRAM	2020
219	32	JAWA BARAT	3203	KABUPATEN CIANJUR	41269972	KILOGRAM	2021

Tampilkan 10 Item dari total 9 — Halaman 1 dari 1

Gambar 6.10 Sampel Data Produksi Kapulaga di Kabupaten Cianjur (Dinas Tanaman Pangan dan Hortikultura Provinsi Jawa Barat, 2022)

Sebelum proses data mining dilaksanakan diawali dengan *data preprocessing* sebagai langkah untuk mempersiapkan data yang bersih dan siap digunakan pada proses selanjutnya. Hasil *data preprocessing* dimasukan kedalam RapidMiner Studio melalui porses *import data*.



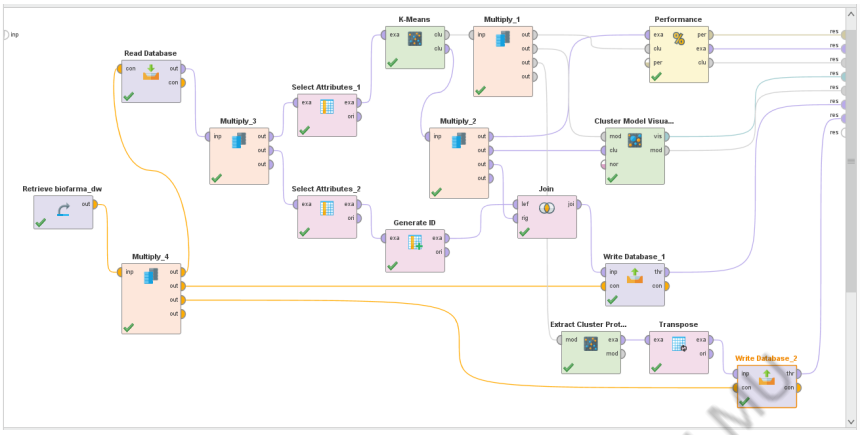
Open in | Turbo Prep | Auto Model — Filter (27 / 27 examples): all

Row No.	Kota	Serai	Jeruk_Nipis	Sambiloto	Lidah_Buaya	Mahkota_De...	Mengkudu	Kapulaga	Temukunci	Temu_Ireng	Temulawak	Lempuyang	Ku
1	KABUPATEN BOGOR	2075340	221000	10128	722266	25812	5911	5591839	5037475	8474	168030	48006	69
2	KABUPATEN SUKABUMI	405500	6000	4405	2092	277424	319302	7613519	35430	13594	12227646	52779	68
3	KABUPATEN CIANJUR	278960	375300	20786	150518	9170838	20242554	106393283	684889	7773	2813553	629235	156
4	KABUPATEN BANDUNG	36915	82000	9662	1658411	166983	99823	551965	39870	29760	1641582	103955	200
5	KABUPATEN GARUT	0	0	0	3270	22440	75572	79030602	23220	897	3797907	7252724	415
6	KABUPATEN TASIKMALAYA	71193	965	277	26788	2572518	2416577	196748021	170	10878	1123586	176543	155
7	KABUPATEN CIAMIS	143950	4415	0	28642	53552	522522	57584013	0	0	179370	3080	286
8	KABUPATEN KUNINGAN	0	1892	0	330	2476	5873	9490029	3390	26162	51600	6531	170
9	KABUPATEN CIREBON	3000	0	1067	0	4620	8211	5783	1495	3585	549	5982	856
10	KABUPATEN MAJALENGKA	0	0	0	200	14539	12770	461364	50610	4835	15175	60	856
11	KABUPATEN SUMEDANG	428800	145000	0	22000	1000	255700	321756	0	0	168	0	111
12	KABUPATEN INDRAMAYU	1868	0	541	1352	1056	8877	0	4580	1335	29131	3192	295
13	KABUPATEN SUBANG	2048500	10500	0	0	104	1857353	208509	515250	0	0	493542	531
14	KABUPATEN PURWAKARTA	2868	3150	9730	29718	400154	785708	9189143	549461	445	345000	53044	928
15	KABUPATEN KARAWANG	6000	290	167108	908	5131910	302158	787018	5103225	768928	8499452	1191309	131
16	KABUPATEN BEKASI	25300	0	0	2500	7500	19459	0	2032100	0	701030	178537	595
17	KABUPATEN BANDUNG BARAT	2095440	266500	9470	63789	236591	203043	5841980	107852	11405	167540	21711	847
18	KABUPATEN PANGANDARAN	0	0	0	2349	249199	20368	24209246	0	0	82975	4369	411

ExampleSet (27 examples, 0 special attr... ...ar attributes) — File Explorer

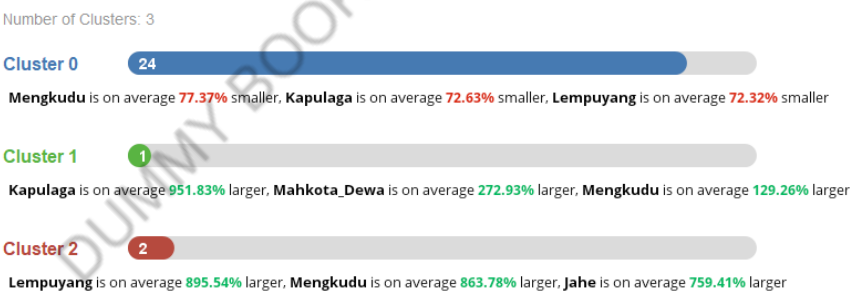
Gambar 6.11 *Import Data* Pada RapidMiner

Pada perancangan proses *data mining* pada RapidMiner Studio menggunakan beberapa operator, diantaranya yaitu *Read/Write Database, Select Attributes, K-Means, Multiply, Cluster Distance Performance, Write Database, Cluster Model Visualizer* dan operator lainnya.



Gambar 6.12 Rancangan *Data Mining*

Proses *K-Means clustering* pada *data mining* dilakukan untuk untuk mengelompokan wilayah potensial penghasil tanaman obat dengan jumlah pengelompokan sebanyak 3 *cluster*. Proses *clustering* menghasilkan pengelompokan wilayah produksi tanaman obat yang tersebar di 3 *cluster*, jumlah wilayah pada masing-masing *cluster* adalah 24 wilayah berada pada *cluster* 0, 1 wilayah berada pada *cluster* 1, dan 2 wilayah berada pada *cluster* 2 dengan nilai *Davies Bouldin Index* sebesar 0,255.



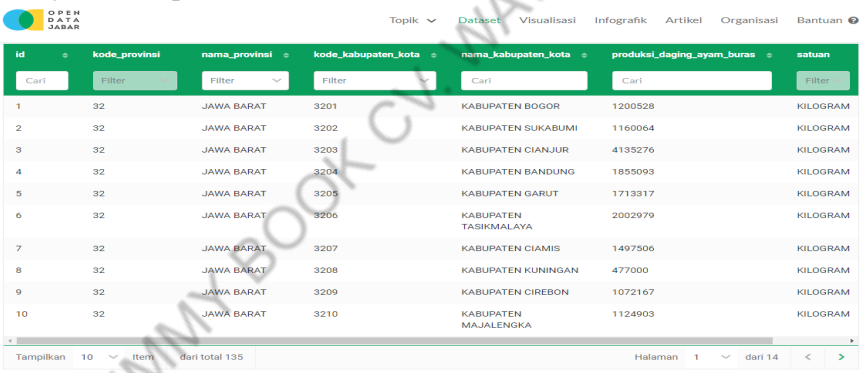
Gambar 6.13 Visualisasi Jumlah Anggota *Cluster*

Data untuk keperluan penelitian diperoleh dari Dinas Ketahanan Pangan dan Peternakan Provinsi Jawa Barat yang dipublikasikan melalui situs web https://opendata.jabarprov.go.id. Informasi tentang data penelitian bisa dilihat di tabel informasi dataset.

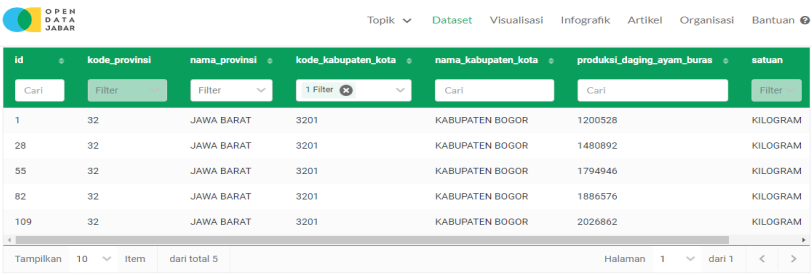
Tabel 6.16 Informasi Dataset Produksi Daging Ayam

No.	Dataset	Dibuat	Diperbarui	Dimensi Awal	Dimensi Akhir	Frekuensi
1	Jumlah Produksi Ayam Buras	03 March 2021 23:54:20	24 May 2022 14:54:07	2016	2020	Tahunan
2	Jumlah Produksi Ayam Pedaging	03 March 2021 23:54:22	24 May 2022 14:39:07	2016	2020	Tahunan
3	Jumlah Produksi Ayam Petelur	03 March 2021 23:54:24	24 May 2022 14:52:07	2016	2020	Tahunan

Dataset yang diperoleh tentang jumlah produksi daging untuk masing-masing jenis ayam terdiri dari delapan atribut yang menjelaskan jumlah produksi daging ayam per tahun di tiap wilayah kabupaten/kota.



Gambar 6.14 Sampel Dataset Produksi Daging Ayam Buras (Dinas Ketahanan Pangan dan Peternakan, 2022)



Gambar 6.15 Sampel Dataset Produksi Daging Ayam Buras di Kabupaten Bogor (Dinas Ketahanan Pangan dan Peternakan, 2022)

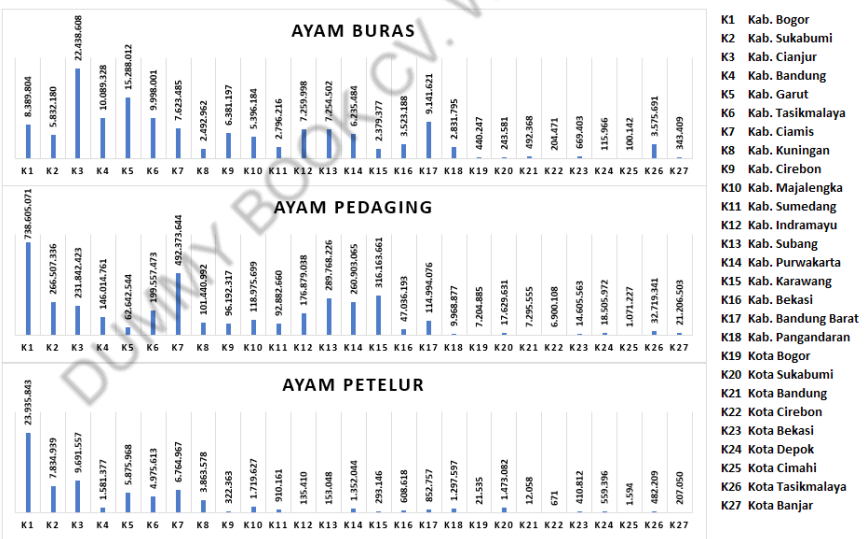
Data preprocessing dilakukan terhadap dataset, kumpulan data diperiksa untuk memastikan data sudah bersih serta melakukan transformasi dan integrasi agar data yang akan digunakan dapat diolah dengan baik pada tahap proses *data mining*. Langkah selanjutnya setelah *data preprocessing* yaitu melakukan *import data* dari format Microsoft Excel ke RapidMiner Studio.

Row No.	nama_kab_kota	buras	pedaging	petelur
1	Kab. Bogor	8389804	738605071	23935843
2	Kab. Sukabumi	5832180	266507336	7834939
3	Kab. Cianjur	22438608	231842423	9691557
4	Kab. Bandung	10089328	146014761	1581377
5	Kab. Garut	15288012	62642544	5875968
6	Kab. Tasikmalaya	9998001	199557473	4975613
7	Kab. Ciamis	7623485	492373644	6764967
8	Kab. Kuningan	2492962	101440992	3863578
9	Kab. Cirebon	6381197	96192317	322363
10	Kab. Majalengka	5396184	118975699	1719627
11	Kab. Sumedang	2796216	92882660	910161
12	Kab. Indramayu	7259998	176879038	135410
13	Kab. Subang	7254502	289768226	153048
14	Kab. Purwakarta	6235484	260903065	1352044

Row No.	nama_kab_kota	buras	pedaging	petelur
15	Kab. Karawang	2379377	316163661	293146
16	Kab. Bekasi	3523188	47036193	608618
17	Kab. Bandung Barat	9141621	114994076	852757
18	Kab. Pangandaran	2831795	9968877	1297597
19	Kota Bogor	440247	7204885	21535
20	Kota Sukabumi	243581	17629631	1473082
21	Kota Bandung	492368	7295555	12058
22	Kota Cirebon	204471	6900108	671
23	Kota Bekasi	669403	14605563	410812
24	Kota Depok	115966	18505972	559396
25	Kota Cimahi	100142	1071227	1594
26	Kota Tasikmalaya	3575691	32719341	482209
27	Kota Banjar	343409	21206503	207050

ExampleSet (27 examples, 0 special attributes, 4 regular attributes)

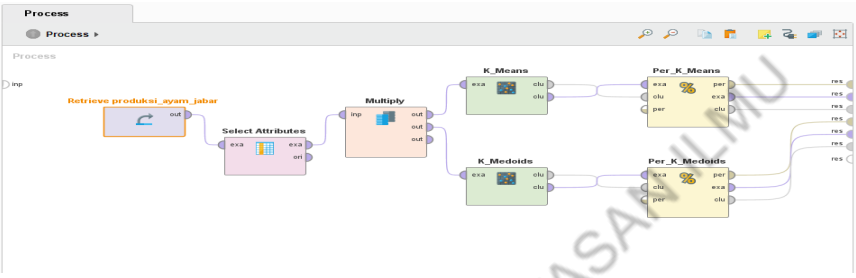
Gambar 6.16 *Import Data* Pada RapidMiner



Gambar 6.17 Visulisasi Data Jumlah Produksi

Pembuatan rancangan proses *data mining* dilakukan sebelum menjalankan proses data lebih lanjut. Beberapa operator untuk pengolahan data digunakan pada RapidMiner Studio, yaitu *Select*

Attributes, *Multiply*, *K-Means*, *K-Medoids* dan *Cluster Distance Performance*. Operator *Select Attributes* digunakan untuk memilih atribut yang akan digunakan pada pengolahan *clustering data* yaitu buras, pedaging dan petelur. *Multiply* berfungsi untuk membuat salinan obyek pada RapidMiner. Operator *k-Means* dan *K-Medoids* adalah operator yang akan menjalankan proses *clustering* sesuai algoritma masing-masing. *Cluster Distance Performance* diperlukan untuk evaluasi *centroid* pada metode *clustering*.



Gambar 6.18 Rancangan RapidMiner Studio

Nilai DBI dari algoritma *clustering K-Means* dan *K-Medoids* sebanyak 6 *cluster* dengan menggunakan RapidMiner dibandingkan untuk mengukur kinerja algoritma terbaik. Perbandingan hasil DBI dapat dilihat pada tabel DBI.

Tabel 6.17 Davies Bouldin Index (DBI)

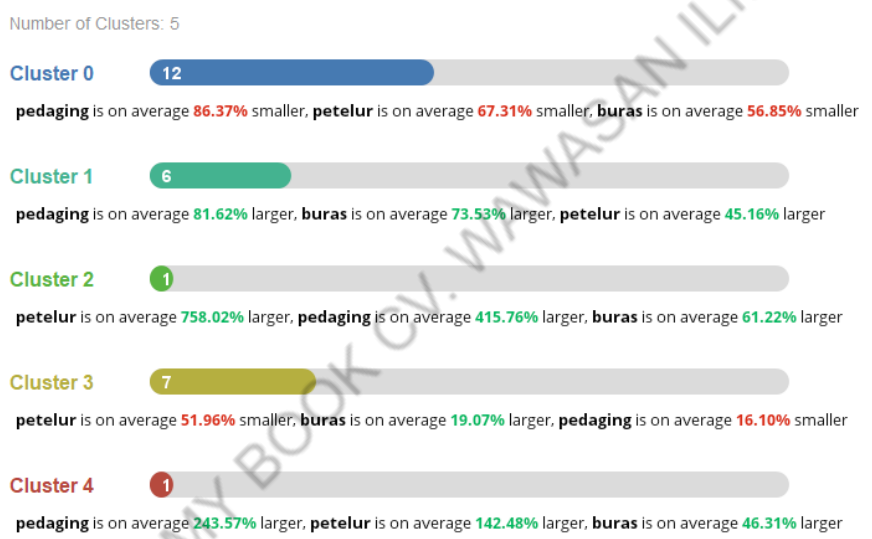
K	K-Means	K-Medoids
2	0,632	0,540
3	0,454	0,338
4	0,411	1,519
5	0,273	0,598
6	0,311	0,918

Dari perbandingan nilai DBI dapat dilihat bahwa nilai DBI terkecil terdapat pada *cluster* 5 menggunakan algoritma *K-Means* yaitu 0,273, hal ini menunjukan kinerja algoritma *clustering* terbaik yang dipilih. Jumlah elemen setiap kelompok hasil proses *clustering* dengan 5 *cluster* menggunakan algoritma *K-Means* dapat dilihat pada tabel jumlah elemen *cluster*.

Tabel 6.18 Jumlah Elemen Cluster

Cluster	Jumlah Elemen
0	12
1	6
2	1
3	7
4	1

Visualisasi data jumlah elemen dan ikhtisar setiap kelompok hasil proses *clustering* dapat dilihat pada grafik jumlah elemen *cluster*.



Gambar 6.19 Grafik Jumlah Elemen *Cluster*

Tujuan dari *K-Means Clustering* adalah untuk menemukan sebuah *centroid* untuk setiap *cluster*, semua objek data kemudian diarahkan ke *centroid* terdekat, yang kemudian membentuk sebuah *cluster*. *Centroid* adalah rata-rata dari semua objek data dalam sebuah *cluster* atau objek data yang paling representatif (Umargono et al., 2020). Nilai *centroid* dari setiap *cluster* yang terbentuk hasil proses data dapat dilihat pada tabel nilai *centroid*.

Tabel 6.19 Nilai Centroid

Cluster	Buras	Pedaging	Petelur
0	2.319.022,75	20.565.533,25	912.549,17
1	9.023.025,33	260.790.364,00	4.050.057,83
2	8.389.804,00	738.605.071,00	23.935.843,00
3	6.222.500,86	121.054.220,40	1.340.753,29
4	7.623.485,00	492.373.644,00	6.764.967,00

Rata-rata *cluster distance* dihitung untuk menilai kinerja algoritma *K-Means* yang digunakan.

```
PerformanceVector:
Avg. within centroid distance: 682499603923448.200
Avg. within centroid distance_cluster_0: 327865706165116.940
Avg. within centroid distance_cluster_1: 1475966683558771.800
Avg. within centroid distance_cluster_2: 0.000
Avg. within centroid distance_cluster_3: 805328675799866.900
Avg. within centroid distance_cluster_4: 0.000
Davies Bouldin: 0.273
```

Gambar 6.20 Rata-rata *Cluster Distance*

Berdasarkan hasil proses *data mining* menunjukan bahwa elemen pada kelompok yang terbentuk sudah terpisah pada masing-masing kelompok, dengan demikian tidak terdapat elemen kelompok yang masuk dalam dua kelompok. Suatu *cluster* dapat dinyatakan konvergen jika tidak terjadi perpindahan atau perubahan elemen dari satu *cluster* ke *cluster* yang lain. Hal ini membuktikan bahwa elemen hasil *clustering* bisa mewakili setiap *cluster*. Nilai pusat *cluster* menunjukan hasil analisis pengelompokan menggunakan algoritma *K-Means* yaitu:

- Cluster 0 berisi produksi ayam buras yang sangat rendah, produksi ayam pedaging yang sangat rendah dan produksi ayam petelur yang sangat rendah.
- Cluster 1 berisi produksi ayam buras yang sangat tinggi, produksi ayam pedaging yang sedang dan produksi ayam petelur yang sedang.
- Cluster 2 berisi produksi ayam buras yang tinggi, produksi ayam pedaging yang sangat tinggi dan produksi ayam petelur yang sangat tinggi.

- Cluster 3 berisi produksi ayam buras yang rendah, produksi ayam pedaging yang rendah dan produksi ayam petelur yang rendah.
- Cluster 4 berisi produksi ayam buras yang sedang, produksi ayam pedaging yang tinggi dan produksi ayam petelur yang tinggi.

Tabel 6.20 *Cluster* Wilayah Produksi Daging Ayam

Cluster 0	Cluster 1	Cluster 2	Cluster 3	
Kab. Garut	Kab. Sukabumi	Kab. Bogor	Kab. Bandung	
Kab. Bekasi	Kab. Cianjur		Kab. Kuningan	
Kab. Pangandaran	Kab. Tasikmalaya		Kab. Cirebon	
Kota. Bogor	Kab. Subang		Kab. Majalengka	
Kota. Sukabumi	Kab. Purwakarta		Kab. Sumedang	
Kota. Bandung	Kab. Karawang		Kab. Indramayu	
Kota. Cirebon			Kab. Bandung Barat	
Kota. Bekasi				
Kota. Depok				
Kota. Cimahi				
Kota. Tasikmalaya				
Kota. Banjar				

Cluster wilayah produksi daging ayam menunjukan pengelompokan wilayah produksi daging ayam di Provinsi Jawa Barat menjadi 5 *cluster*. *Cluster* 0 merupakan *cluster* yang memiliki anggota kelompok paling banyak dibandingkan dengan *cluster* yang lain. Hasil *clustering* dapat dimanfaatkan dalam proses bisnis terkait informasi jumlah produksi daging ayam di wilayah Jawa Barat sebagai acuan dalam pola pembinaan untuk meningkatkan

produksi pangan hewani, pengembangan potensi budidaya ayam, dan pengembangan potensi distribusi pakan ternak.

6.4 Latihan Soal

1. Bank TI ingin merumuskan strategi pemasarannya kembali. Hal pertama yang ingin dilakukan adalah segmentasi pelanggannya. Data https://docs.google.com/spreadsheets/d/1EoM9UaQy_ccSCcwEd16cRz4iTxxk_PhO/edit?usp=sharing&ouid=107032425374168372161&rtpof=true&sd=true memuat 9000 pengguna aktif kartu kredit Bank TI selama 6 bulan. Segmentasi pasar dilakukan berdasarkan variabel perilaku pengguna kartu kredit. Terdapat 17 variabel perilaku pengguna yang dapat digunakan untuk segmentasi pasar. Lakukan analisis kluster untuk membantu Bank TI melakukan segmentasi pelanggannya dan berikan penjelasan kecenderungan perilaku pelanggan di tiap segmen yang terbentuk! Diperbolehkan menggunakan bantuan sortware R, Weka maupun Knime

   Keterangan Variabel:

CUSTID	Identification of Credit Card holder (Categorical)
BALANCE	Balance amount left in their account to make purchases
BALANCE FREQUENCY	How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
PURCHASES	Amount of purchases made from account
ONE OFF PURCHASES	Maximum purchase amount done in one-go
INSTALLMENTS PURCHASES	Amount of purchase done in installment
CASH ADVANCE	Cash in advance given by the user

PURCHASES FREQUENCY	How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
ONE OFF PURCHASES FREQUENCY	How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASES INSTALLMENTS FREQUENCY	How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
CASH ADVANCE FREQUENCY	How frequently the cash in advance being paid
CASH ADVANCE TRX	Number of Transactions made with "Cash in Advanced"
PURCHASES TRX	Number of purchase transactions made
CREDIT LIMIT	Limit of Credit Card for user
PAYMENTS	Amount of Payment done by user
MINIMUM_ PAYMENTS	Minimum amount of payments made by user
PRC FULL PAYMENT	Percent of full payment paid by user
TENURE	Tenure of credit card service for user

2. Ditentukan banyaknya cluster yang dibentuk dua **(k=2)**. Banyaknya cluster harus lebih kecil dari pada banyaknya data **(k<n)**. Inisialisasi centroid dataset pada tabel dataset diatas adalah **C1 = {1 , 1}** dan **C2 = {2 , 1}.** Hitunglah jarak data dengan **Centroid** dengan menggunakan K Means Clustering dari data berikut ini.

n	a	b
1	1	1
2	2	1
3	4	3
4	5	4

3. Cari data Corona di website https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=-COVID19_line_list_data.csv di Indonesia. Kemudian buatlah Analisa clusteringnya dengan menggunakan software dan jelaskan analisanya.

BAB VII
DATA MINING DAN ERP (ENTERPRISE RESOURCE PLANNING) DENGAN MOONSON ACADEMY

7.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk mengenalkan penggunaan data mining dan ERP (*Enterprise Resource Planning*) dengan aplikasi Moonson Academy.

Materi yang diberikan :

1. Konsep Data Mining dan ERP
2. Penjelasan aplikasi Moonson Academy
3. Data Mining dan ERP dengan Moonson Academy

Capaian Pembelajaran:

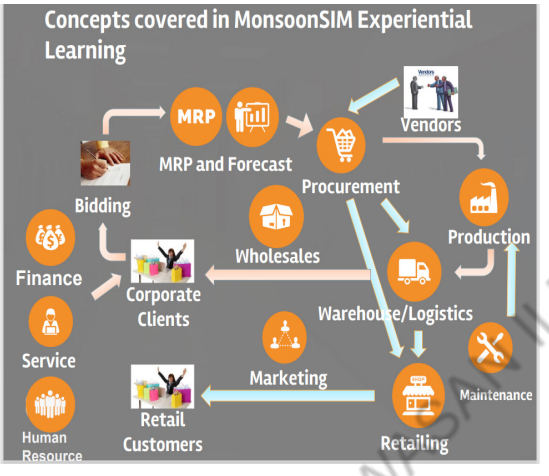
Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menggunakkan aplikasi Moonson Academy utuk data mining dan ERP (*Enterprise Resource Planning*). Capaian pembelajaran lulusan adalah mahasiswa mampu.

7.2 Pendahuluan

Penelitian menunjukkan bagaimana kita mendapatkan pengetahuan melalui model 70:20:10 untuk pembelajaran dan pengembangan dikembangkan dari penelitian di Center for Creative Leadership (CCL) di North Carolina pada tahun 1980. *MonsoonSim Experiential Learning* adalah pembelajaran baru berdasarkan berbasis cloud, business process experiential dan platform pembelajaran.

7.3 Data Mining dan ERP dengan Monsoon Academy

Konsep yang terlingkup dalam *MonsoonSim Experiential Learning* meliputi konsumen retail, konsumen perusahaan



Gambar 7.1 Konsep MonsoonSIM Academy

Bisnis konsep meliputi keuangan, logistic, manufaktur, *business intelligence*, perawatan, pergudangan, transportasi, *outsourcing*, sumber daya manusia, operasi binis keseluruhan.

Perbedaan yang didapatkan oleh mahasiswa dalam pembelajaran tradisional dengan pembelajaran menggunakan MonsonSim adalah :

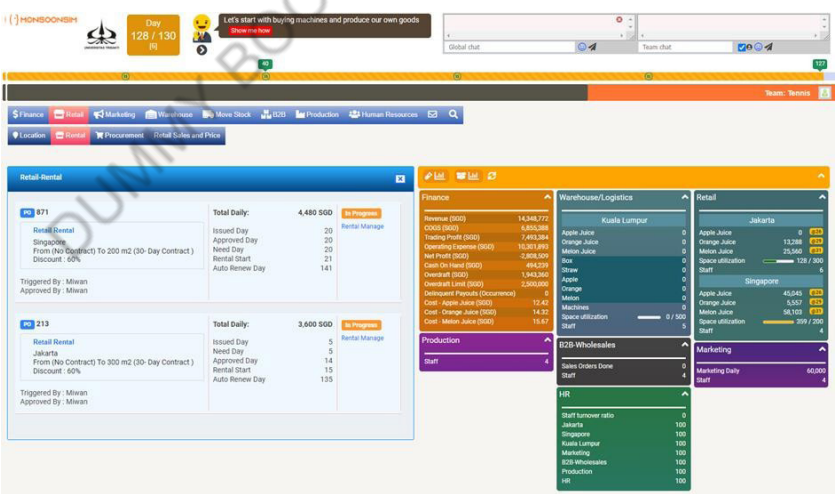
Tradisional	Pembelajaran Menggunakan MonsoonSIM
Pembelajaran dengan teori	Pembelajaran sambal mengerjakan
Lingkungan yang tidak kompetitif	Lingkungan yang kompetitif
Pembelajaran individu	Tim kolaborasi dan brainstorming
Tidak ada kesalahan yang dibuat	Kesalahan dibuat dan dipelajari
Tidak ada feedback	Feedback yang cepat

1. Dashboard

Gambar 7.2 Dashboard

Gambar 7.2 merupakan gambar Dashboard pada monsonsim

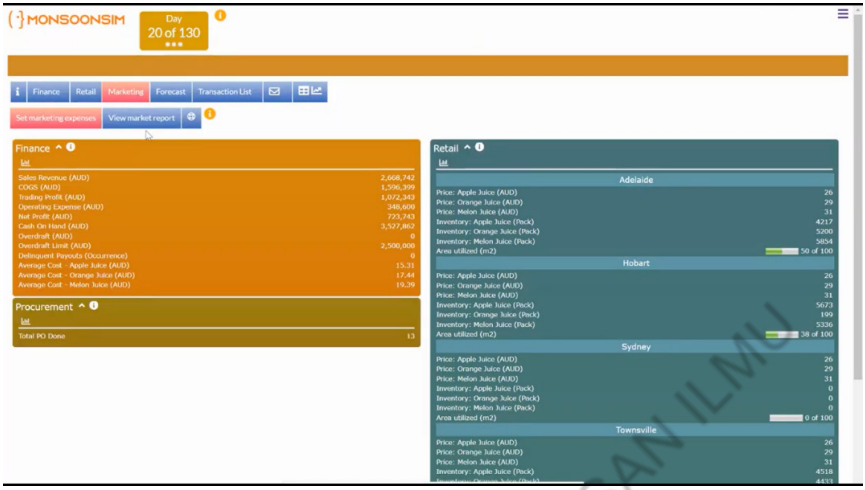
2. Retail



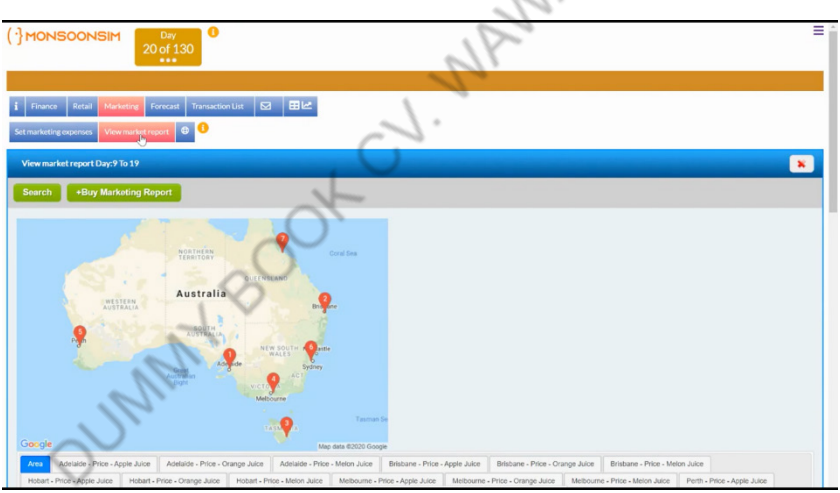
Gambar 7.3 Retail

Gambar 7.3 merupakan Gambar Retail pada monsoonsim

BASIC RETAIL STRATEGY

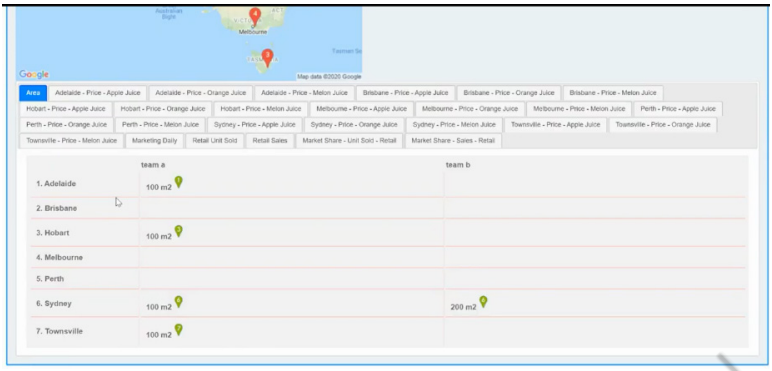


Gambar 7.4 Retail Tim A



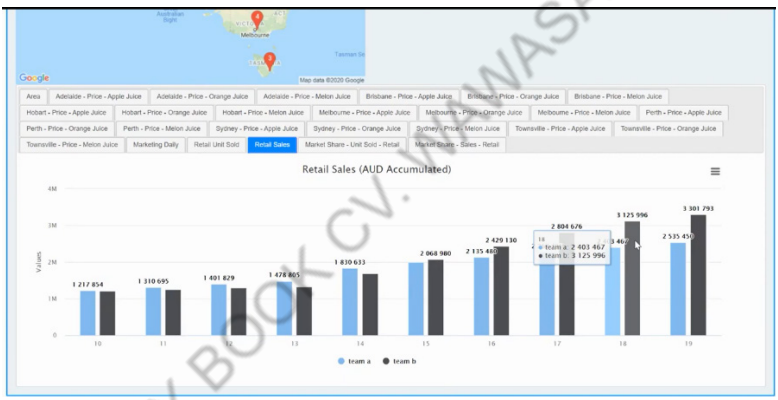
Gambar 7.5 Marketing Report

Tim A memiliki 4 toko retail di Adelaide, Hobart, Sydney dan Townsville. Kemudian tim A membeli marketing report untuk mengetahui keadaan tim A disbanding kompetitornya.



Gambar 7.6 Lokasi Team A dan Team B

Kemudian klik tab area, setelah dilihat competitor tim B hanya memiliki 1 toko di Sydney. Kemudian klik Retail Sales.



Gambar. 7.7 Retail Sales

Gambar. 7.8 Market Share Retail

Dari Retail Sales, diketahui penjualan tim B lebih tinggi dibandingkan tim A yang memiliki 4 toko. Pada tab market share-sales-retail terlihat juga persentase akumulasi tim B sebesar 56,6%, lebih tinggi dibandingkan tim A yaitu sebesar 43,4%.

Yang bisa kita Analisa dengan cara sebagai berikut :

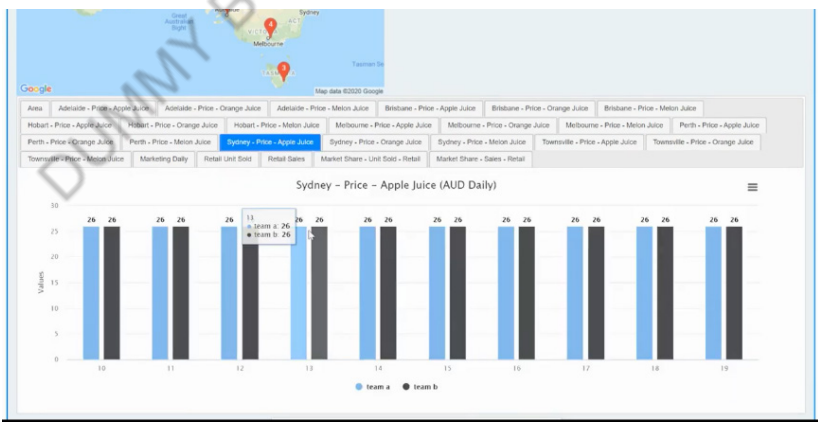
1. Mengecek forecast produk orange juice setiap toko

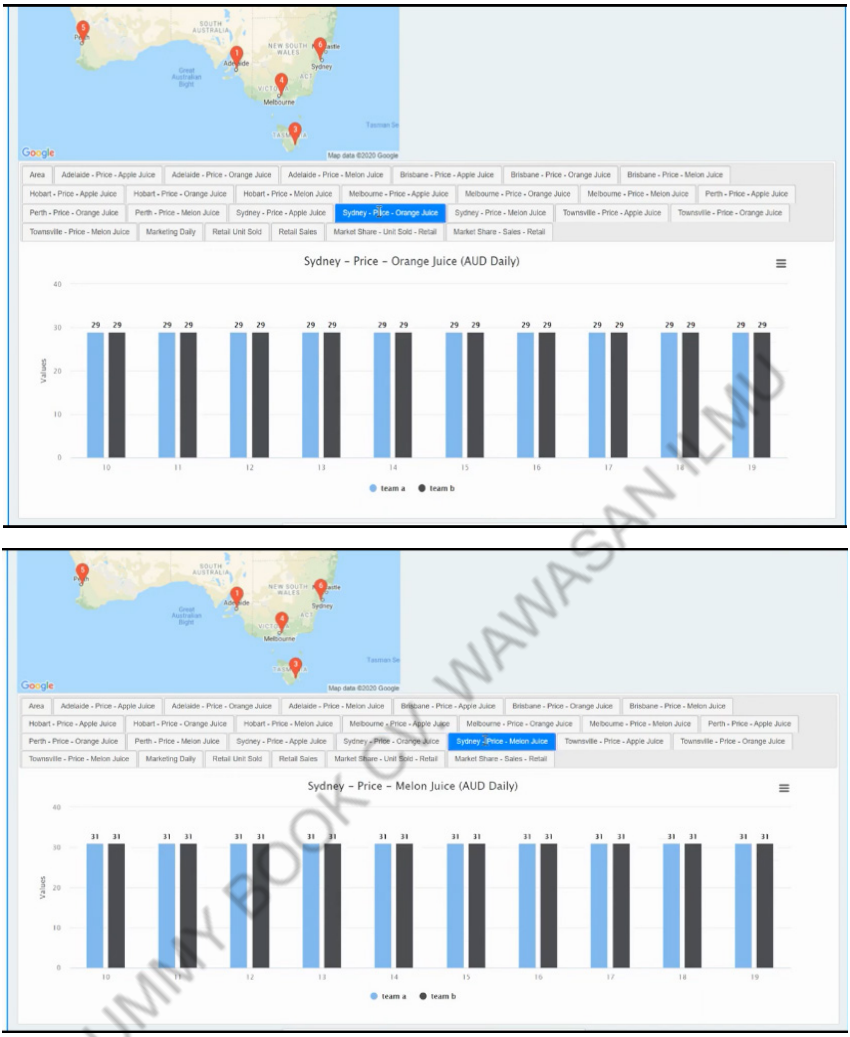
Jika dijumlahkan setiap produk orange juice untuk 3 toko (Hobart, Adelaide, Townsville) didapatkan total 279 unit, berbanding jauh dengan *demand* yang ada di Sydney yaitu sebesar 485 unit.

Dapat diasumsikan bahwa Sydney merupakan pasar yang paling penting.

Gambar. 7.9 Retail Demand

2. Membandingkan harga orange juice, apple juice dan melon juice di Sydney

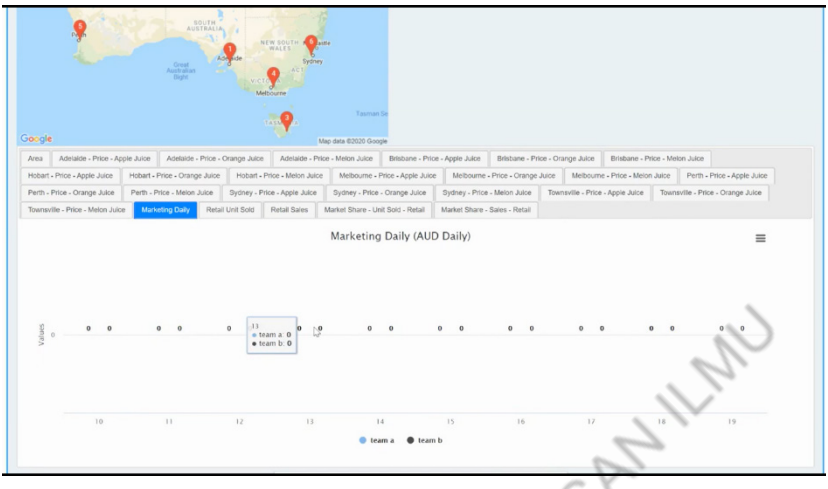




Gambar. 7.10 Harga Apple Jus, Orange Jus, Melon Jus dan di Sidney

Dari ketiga produk tersebut, terlihat bahwa ketiga produk memiliki harga yang sama pada Tim A maupun Tim B. Jadi perbedaan Retail Sales tadi bukan karena harga.

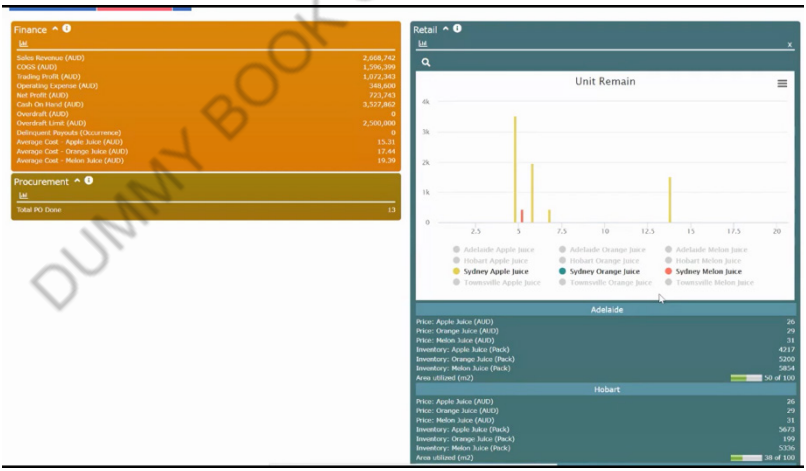
3. Mengecek Marketing



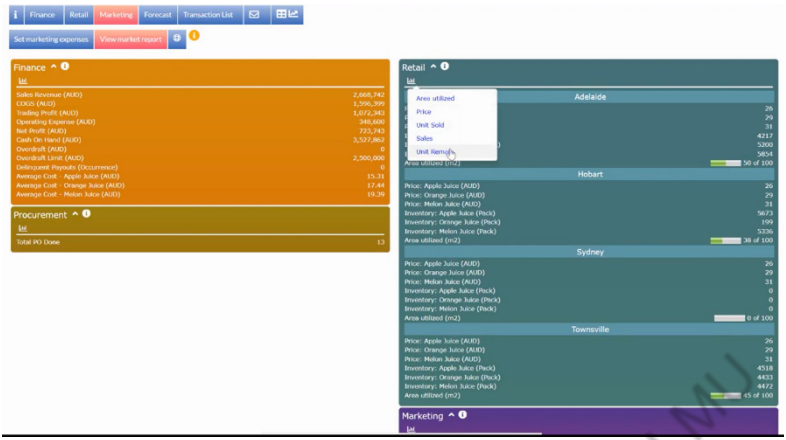
Gambar 7.11 Marketing Daily

Pada Gambar terlihat kedua tim tidak memasang marketing, bisa dipastikan bahwa perbedaan retail sales bukan karena marketing.

4. Mengecek Unit Remain



Gambar 7.12 Unit Remain

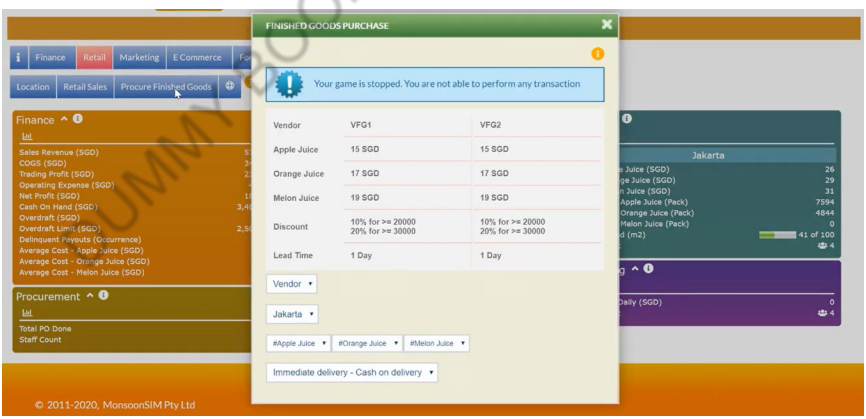


Gambar 7.13 Pengechekan Unit Remain

Tim A sering kehabisan barang di Sydney. Maka dari itu bisa disimpulkan, bahwa penjualan retail Tim A lebih sedikit dibandingkan Tim B karena sering kehabisan barang di Sydney, di mana Sydney merupakan area penting karena memiliki *demand* paling tinggi di antara toko-toko.

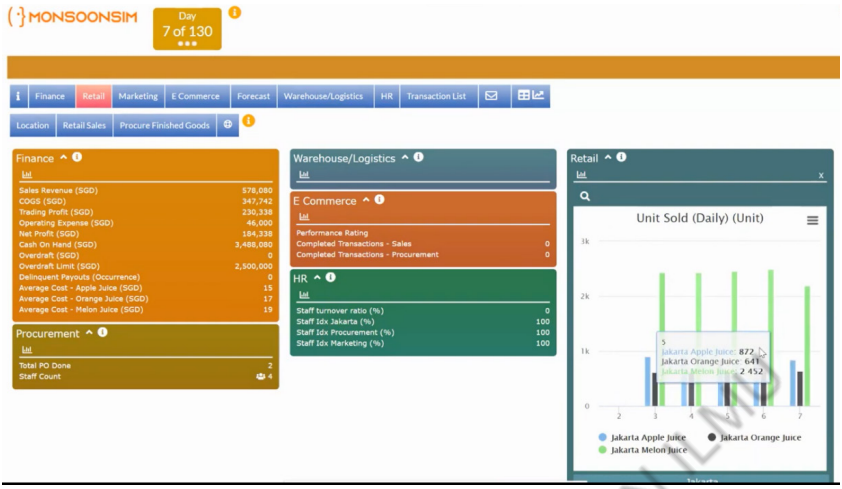
FITUR PEMBELIAN (PROCUREMENT)

Ketika bermain Moonson Sim, pasti ada fiture *Finished Goods Purchase*.



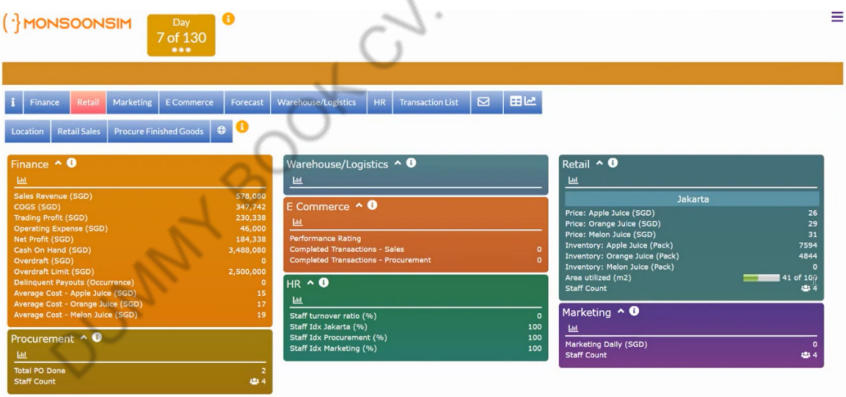
Gambar 7.14 Finished Good

Yang harus kita perhatikan ketika membeli barang ialah Lead Time dari masing-masing vendor dan melihat trend penjualan beberapa hari.



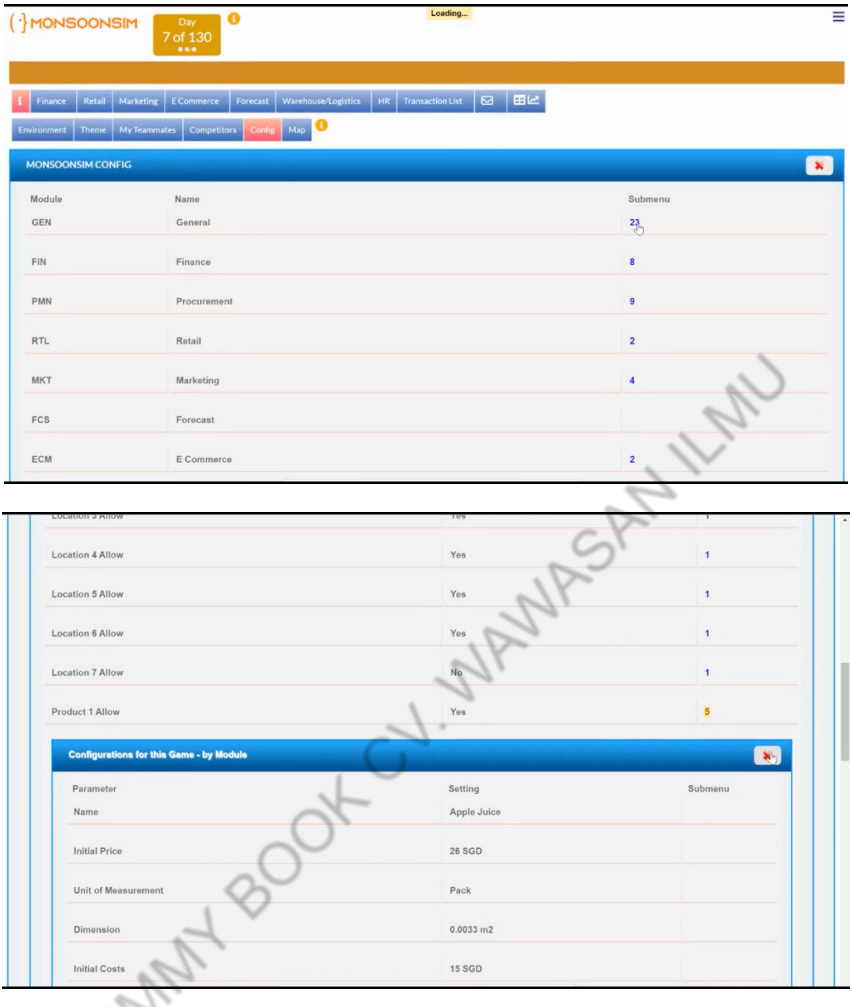
Gambar 7.15 Unit Sold (Daily)

Jika Vendor membutuhkan waktu mengirim 1 hari, ketika sisa stok barang sudah sekitaran 1 hari lebih punya stok. Untuk mengetahui banyak yang harus kita beli, kita dapat melihat informasi luas toko dan dimensi produk.



Gambar 7.16 Inventory

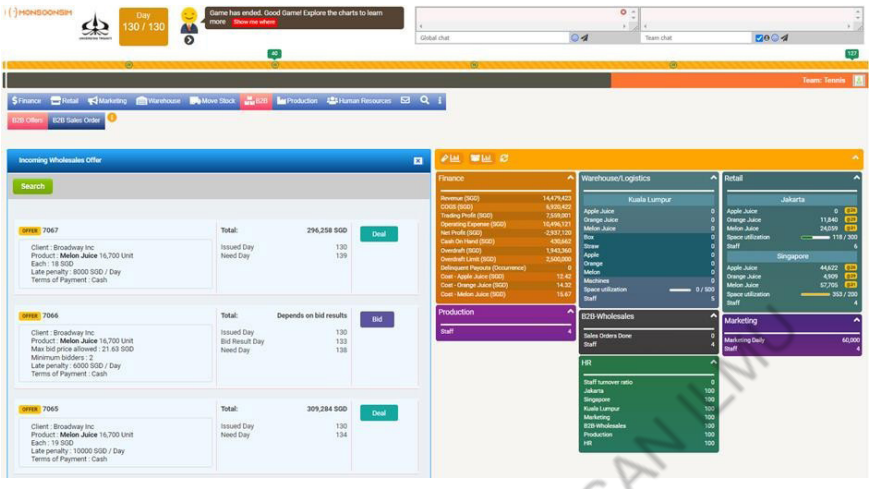
Dapat diketahui bahwa area toko sebesar 100 . Kemudian untuk dimensi produk dapat dilihat di bagian config dan klik bagian submenu general



Gambar 7.17 Konfigurasi

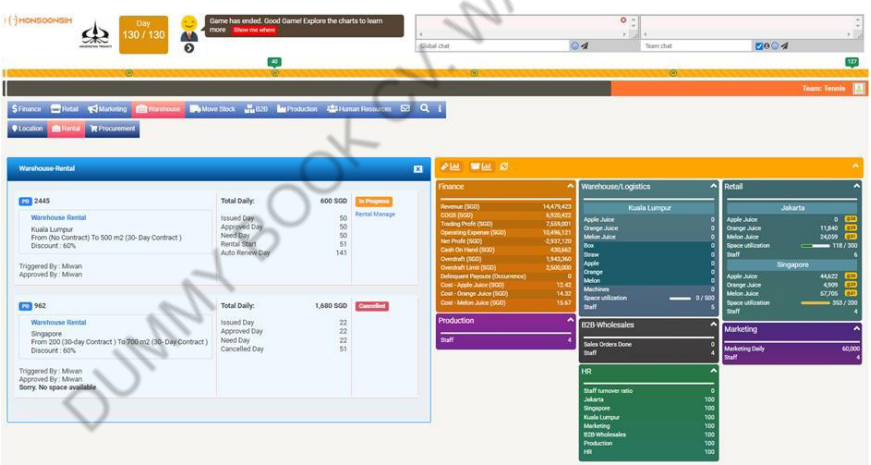
Terlihat dimensi untuk produk *Apple Juice* sebesar 0,0033 . Dengan besar lokasi sekitar 100 dengan dimensi produk 33, maka total maximum produk di toko sebanyak 30.000 unit.

3. B2B



Gambar 7.18 B2B

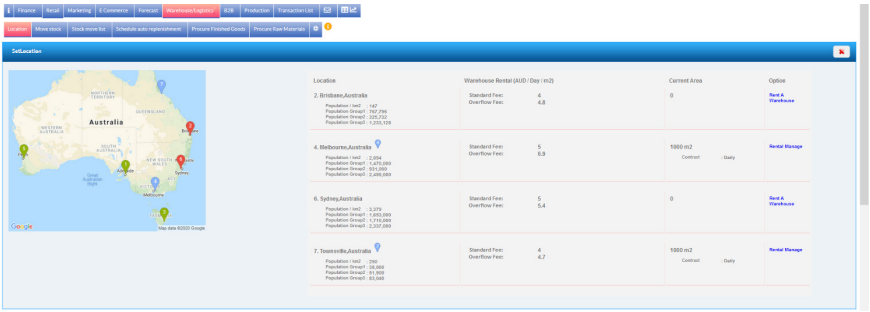
4. Warehouse



Gambar 7.19 Warehouse

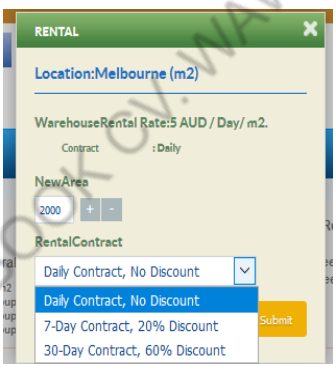
Bagaimana mengatur penyewaan Warehouse

1. Mengklik bagian "*Warehouse/Logistics*" tab untuk menampilkan "*Location*" sub tab. Kemudian mengklik "*Location*" sub tab.
2. Mengklik "*Rental Manage*" untuk lokasi yang diinginkan.



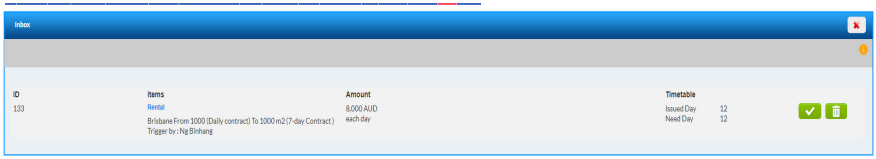
Gambar 7.20 Warehouse

3. "Rental" akan muncul, dan dapat mengatur area warehouse serta memilih kontrak sewa yang disukai.
4. Pengguna dapat menambahkan atau mengurangi area warehouse dan dapat memilih tipe kontrak rental.
5. Setelah memastikan area dan kontrak yang tepat, pengguna dapat klik "submit" untuk membuat PO.

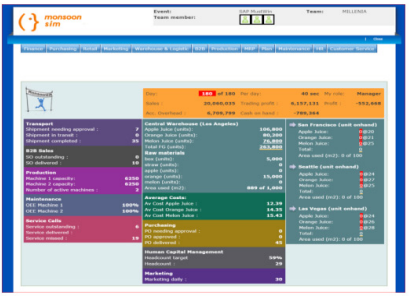


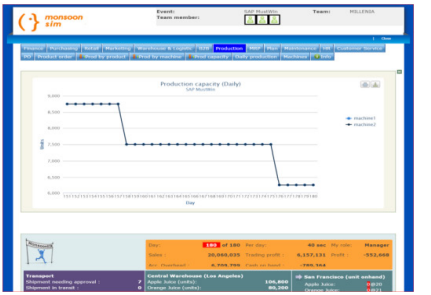
Gambar 7.21 Warehouse Rental

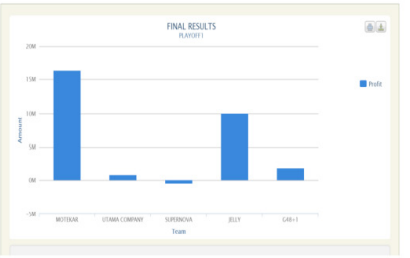
6. Sebuah PO akan dihasilkan, pengguna dapat memilih apakah akan menyetujui atau membatalkan PO untuk penyesuaian sewa warehouse.

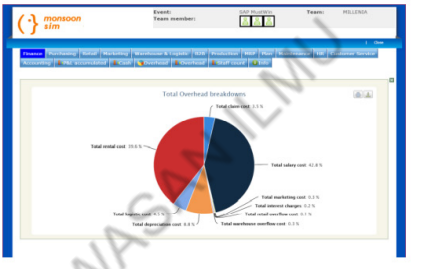


Gambar 7.22 PO Rental









Gambar 7.23 Grafik Penjualan

Konsep bisnis yang kompleks dibuat menjadi mudah

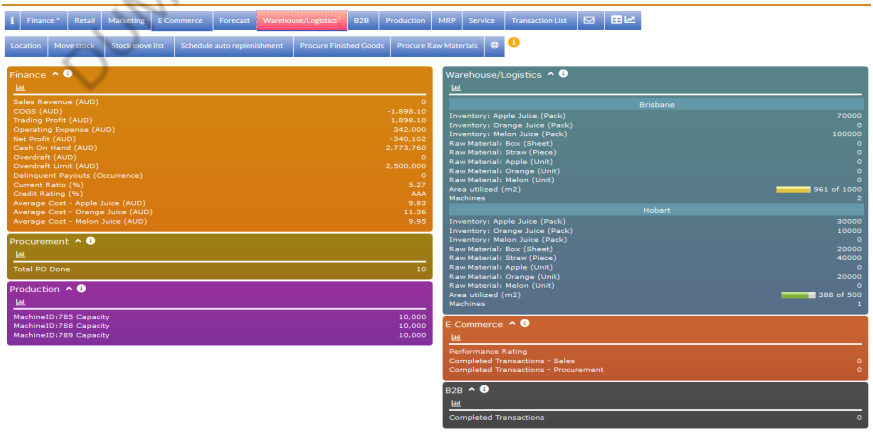
Mengecek *Bill of Material*

1. Pengguna dapat klik tab "i" untuk menampilkan sub tab "config", kemudian klik subtab "config" untuk menampilkan "MoonsoonSIM Config".
2. Pengguna dapat mencari "PRD" dan klik nomor yang ada disebelahnya.
3. Pengguna dapat melihat *Bill of Material* setiap produk.

Gambar 7.24 Gambar Bill of Material

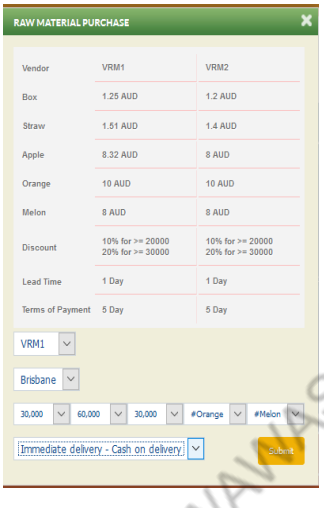
Bagaimana membeli bahan baku

1. Pengguna dapat klik "Warehouse/Logistic" tab untuk menunjukkan "Procure Raw Material" sub tab.
2. Pengguna dapat klk "Procure Raw Materials" sub tab untuk membeli bahan baku.



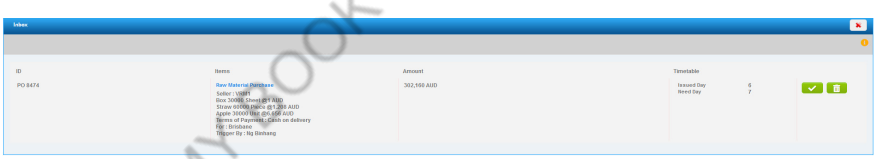
Gambar 7.25 Procure raw material di Warehouse

3. Pop out "Material Purchase" akan muncul. Kemudian pengguna akan memutuskan vendor yang disukai, lokasi dan jumlah bahan baku. Kemudian, pengguna mengklik "Submit" untuk menghasilkan PO.



Gambar 7.26 Raw Material Purchase

4. PO akan dihasilkan, kemudian pengguna dapat memutuskan apakah akan menyetujui PO atau menolaknya.



Gambar 7.27 Purchase Order

MONITORING PRODUCTION CAPACITY

1. Pengguna klik "Production" tab untuk menunjukkan "Production" sub tab.
2. Pengguna kemudian klik "Production" sub tab untuk melihat grafik kapasitas produksi secara keseluruhan.
3. Pengguna dapat memantau disini utuk mengetahui kapan produk telah diproduksi dan berapa banyak kuantitas yang telah diproduksi juga.

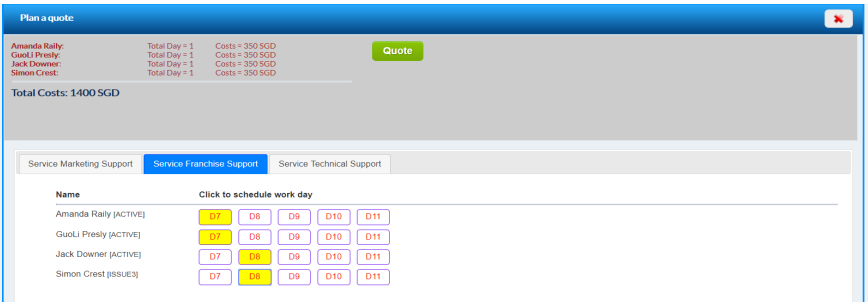
Gambar 7.28 Kapasitas Produksi

How to do Service Sales

1. Pengguna dapat klik pada "Service" tab, kemudian klik "Incoming Service Request".
2. Daftar proposal permintaan layanan akan ditampilkan. Pengguna dapat menganalisis dukungan yang dibutuhkan dan memilih proposal permintaan layanan terbaik.
3. Pengguna dapat mengklik "Plan a quote" untuk mengutip harga untuk permintaan layanan.

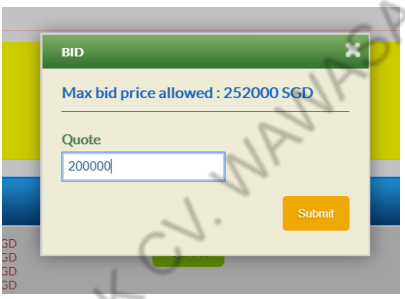
Gambar 7.29 Plan a quote

4. Pengguna dapat memilih dukungan yang diperlukan untuk permintaan layanan.
5. Pengguna kemudian dapat mengatur jadwal untuk setiap staf yang ditugaskan untuk permintaan layanan.
6. Pengguna dapat mengklik "Quote" setelah selesai menjadwalkan staf.



Gambar 7.30 Service Francise Support

7. Pengguna dapat menempatkan kutipan yang diinginkannya.
8. Pengguna dapat mengklik "Submit" untuk mengirimkan penawaran untuk permintaan layanan.



Gambar 7.31 Penawaran

BAB VIII
BIG DATA, HADOOP DAN MAP REDUCE

8.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk mengenalkan pengertian mengenai Big Data, Hadoop dan Map Reduce.

Materi yang diberikan :

1. Pengertian Big Data
2. Hadoop
3. Map Reduce

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai pemahaman mengenai Big Data, Hadoop dan Map Reduce. Capaian pembelajaran lulusan adalah mahasiswa mampu.

8.2 Pendahuluan

Perkembangan teknologi informasi dan komunikasi juga membawa dampak yang besar bagi perkembangan berbagai sumber dan variasi tipe data khususnya di bidang industri manufaktur dan jasa. Data yang dimaksud tidak lagi hanya data terstruktur yang diperoleh dari hasil ekperimen maupun percobaan, atau dari hasil penyebaran kuesioner serta laporan harian perusahaan. Data secara tradisional masih dikenal hanya berupa angka yang menggambarkan jumlah, kinerja atau skala. Pada kenyataannya saat ini banyak data yang tidak terstruktur berasal dari sosial media seperti twitter, instagram, facebook maupun website, review produk di e-commerce atau platform pribadi perusahaan, dan juga dapat berupa gambar produk itu sendiri. Selanjutnya kumpulan data terstruktur dan tidak terstruktur lebih dikenal dengan *big data* atau data yang besar.

Data terstruktur memerlukan data yang dikategorikan dan disimpan dalam file menurut deskripsi format tertentu, di mana data tidak terstruktur adalah teks bentuk bebas seperti contoh yang telah disebutkan. Contoh sederhana adalah ponsel dari masa lalu telah berevolusi menjadi smartphone mampu mengirim pesan teks, sebagai mesin pencarian, menelepon, dan memainkan sejumlah perangkat lunak berbasis aplikasi. Semua aktivitas yang dilakukan di ponsel ini dapat menghasilkan aset data yang dapat dilacak. Komputer dan tablet terhubung ke platform terkait Internet (media sosial, aktivitas situs web, iklan melalui platform video) semuanya menghasilkan data. Teknologi pemindaian dapat membaca konsumsi energi, elemen terkait perawatan kesehatan, aktivitas lalu lintas, bahkan yang terbaru adalah dapat melacak mobilitas masyarakat selama pandemi dan masih banyak lainnya. Banyak data tidak terstruktur yang belum banyak dimanfaatkan untuk menggali informasi penting yang terkandung di dalamnya. Bab ini akan mengenalkan tentang big data dan konsep data mining.

Big Data" mirip dengan "data kecil", tetapi ukurannya lebih besar. Tetapi memiliki data yang lebih besar itu membutuhkan pendekatan, Teknik, alat, dan arsitektur yang berbeda Tujuan untuk memecahkan masalah lama atau baru dengan cara yang lebih baik. Big Data menghasilkan nilai dari penyimpanan dan pemrosesan informasi digital dalam jumlah yang sangat besar yang tidak dapat dianalisis dengan teknik komputasi tradisional. Big data berbeda karena secara otomatis digenerate oleh mesin (misalnya Sensor tertanam dalam mesin)

1. Biasanya sumber yang sama sekali baru dari data. (Misalnya penggunaan internet)
2. Tidak dirancang untuk bersikap ramah. (misalnya teks streaming).
3. Mungkin tidak memiliki nilai, perlu fokus pada bagian yang penting.

Fenomena analitik data besar terus-menerus tumbuh sebagai organisasi merombak proses operasional mereka mengandalkan data langsung dengan harapan dapat mendorong pemasaran yang efektif teknik, meningkatkan keterlibatan pelanggan, dan berpotensi menyediakan produk dan layanan baru (Zhuge, 2015) (SaS, 2015)

1) Apa itu Big Data?

Big Data mengacu pada kumpulan besar data kompleks, keduanya terstruktur dan tidak terstruktur yang pengolahan tradisional teknik dan/atau algoritma tidak dapat dioperasikan. Ini bertujuan untuk mengungkapkan pola tersembunyi dan telah menyebabkan evolusi dari paradigma sains berbasis model menjadi sains berbasis data paradigma. Menurut sebuah studi oleh Boyd & Crawford (Boyd dan Crawford, 2011) itu "bertumpu pada interaksi dari:

a) Teknologi: memaksimalkan daya komputasi dan akurasi algoritmik untuk mengumpulkan, menganalisis, menghubungkan, dan membandingkan kumpulan data yang besar.
b) Analisis: menggambar pada kumpulan data besar untuk mengidentifikasi pola untuk membuat ekonomi, sosial, teknis, dan tuntutan hukum.
c) Mitologi: kepercayaan luas bahwa kumpulan data besar menawarkan bentuk kecerdasan dan pengetahuan yang lebih tinggi yang dapat menghasilkan wawasan yang sebelumnya tidak mungkin, dengan aura kebenaran, objektivitas, dan akurasi."

Ilmuwan IBM menyebutkan bahwa Big Data memiliki empat dimensi: Volume, Kecepatan, Ragam, dan Kebenaran (IBM Big Data & Analytics Hub. 2014). Gartner setuju dengan IBM dengan menyatakan bahwa Big Data terdiri dari "volume tinggi, Deskripsi empat dimensi dirinci di bawah ini (Katal, 2013):

a) Volume – Data saat ini yang ada dalam petabyte, yang sudah bermasalah; diprediksi bahwa dalam beberapa tahun ke depan akan meningkat menjadi zettabytes (ZB) (Katal, 2013). Ini karena meningkatnya penggunaan ponsel perangkat dan jaringan sosial terutama.
b) Velocity – Mengacu pada kecepatan data ditangkap dan kecepatan aliran data. Ditingkatkan ketergantungan pada data langsung menyebabkan tantangan bagi analitik tradisional karena datanya terlalu besar dan terus menerus bergerak.
c) Variasi – Karena data yang dikumpulkan tidak spesifik kategori atau dari satu sumber, ada berbagai format data mentah, diperoleh dari web, teks, sensor, email, dll. Yang

terstruktur atau tidak terstruktur. Jumlah yang besar ini menyebabkan metode analitis tradisional lama gagal dalam mengelola data besar.

d) Kebenaran – Ambiguitas dalam data adalah yang utama fokus dalam dimensi ini – biasanya dari kebisingan dan kelainan dalam data.

Melengkapi perusahaan dengan e-commerce berbasis Big Data arsitektur membantu dalam mendapatkan "wawasan pelanggan" yang luas perilaku, tren industri, keputusan yang lebih akurat untuk meningkatkan hampir setiap aspek bisnis, dari pemasaran dan periklanan, hingga merchandising, operasi, dan bahkan retensi pelanggan"(The Big Data Landscape, 2014). Namun data yang sangat besar diperoleh dapat menantang Empat V yang disebutkan sebelumnya. Mengacu pada contoh berikut, wawasan tentang bagaimana Big Data membawa nilai bagi organisasi disebutkan.

United Parcel Service (UPS) memahami pentingnya besar analitik data di awal tahun 2009 yang mengarah pada pemasangan sensor dilebih dari 46.000 kendaraannya; ide di balik ini adalah untuk mencapai "kecepatan dan lokasi truk, berapa kali itu ditempatkan secara terbalik dan apakah sabuk pengaman pengemudi adalah tertekuk. Sebagian besar informasi diunggah di akhir hari ke pusat data UPS dan dianalisis dalam semalam" (Rosenbush, 2013). Oleh menganalisis sensor efisiensi bahan bakar dan data GPS, UPS mampu mengurangi konsumsi bahan bakar sebesar 8,4 juta galon dan mengurangi durasi rutenya menjadi 85 juta mil (Rosenbush, 2013).

Andrew Pole, analis data untuk pengecer Amerika Target Corporation, mengembangkan model prediksi kehamilan, sebagai ditunjukkan dengan nama pelanggan diberi " prediksi kehamilan" (Charles Duhigg. 2012) skor. Selanjutnya Target bisa mendapatkan wawasan tentang bagaimana 'hamil seorang wanita'. Tiang awalnya digunakan Baby-shower-registri Target untuk mendapatkan wawasan tentang wanita kebiasaan berbelanja, membedakan kebiasaan itu saat mereka mendekat tanggal jatuh tempo mereka; yang "perempuan didaftar itu dengan sukarela" diungkapkan" (Charles Duhigg. 2012). Setelah tahap pengumpulan data awal, Pole dan timnya menjalankan serangkaian tes, menganalisis kumpulan data dan menyimpulkan secara efektif pola-pola yang dapat berguna bagi perusahaan.

2) Mengapa transformasi dari analitik tradisional ke Analisis Big Data diperlukan?
3) Bagaimana memenuhi permintaan Sumber Daya Komputasi? (Bagian II)
4) Apa implikasi Big Data terhadap evolusi? Penyimpanan Data? (Bagian III)
5) Apa saja inkonsistensi Big Data? (Bagian IV)
6) Bagaimana Big Data dipetakan ke dalam ruang pengetahuan? (Bagian V)

8.3 Pendalaman Big Data

Big Data merupakan istilah yang berlaku untuk informasi yang tidak dapat diproses atau dianalisis menggunakan alat tradisional. Data yang melebihi proses kapasitas dari sistem database konvensional yang ada juga dikenal dengan Big Data. Data juga memiliki ukuran terlalu besar dan terlalu cepat bertambah maupun berubah atau tidak sesuai dengan struktur arsitektur database yang ada. Penggalian informasi dalam data jenis ini harus dilakukan dengan alternatif metode dan alat yang tepat.

Sebelum era big data dimulai, sektor bisnis dan ekonomi memberikan nilai yang relatif rendah terhadap data yang mereka kumpulkan yang tidak memiliki nilai langsung. Ketika era big data dimulai, investasi dalam pengumpulan dan penyimpanan ini dilakukan untuk melihat potensi nilai data di masa depan berubah, dan organisasi membuat upaya untuk menyimpan setiap bit data yang potensial untuk perkembangan bisnisnya. Pergeseran perilaku ini menciptakan lingkaran yang baik di mana data disimpan dan kemudian, karena data tersedia, orang-orang ditugaskan untuk menemukan nilai di dalamnya bagi organisasi. Keberhasilan dalam menemukan nilai menyebabkan lebih banyak data yang dikumpulkan dan segera. Beberapa data yang disimpan adalah jalan buntu, tetapi sering kali hasilnya dikonfirmasi bahwa semakin banyak data yang Anda miliki, semakin baik bagi Anda informasi yang dapat dimanfaatkan. Perubahan besar lainnya diawal era big data adalah perkembangan yang pesat, penciptaan, dan kematangan teknologi untuk menyimpan, memanipulasi, dan menganalisis data

ini dengan cara yang baru dan lebih efisien. Sekarang kita berada diera big data, tantangan kita bukanlah datanya tetapi mendapatkan data yang benar dan menggunakan komputer untuk meningkatkan pengetahuan dan mengidentifikasi pola yang tidak kita lihat atau tidak bisa ditemukan sebelumnya.

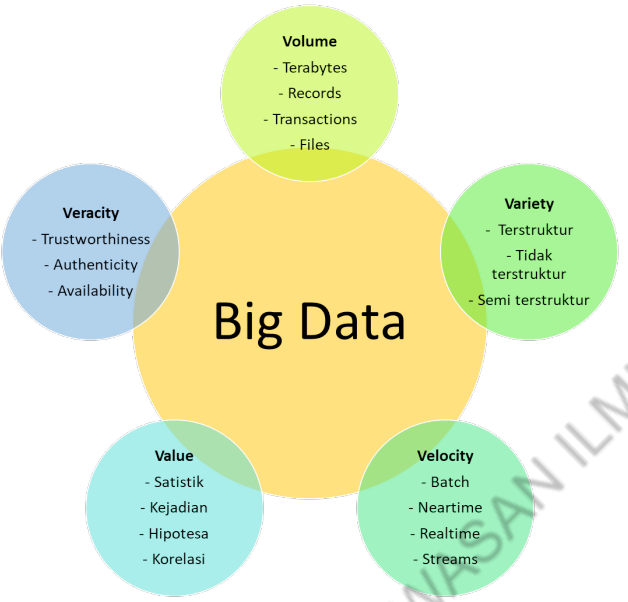
Transformasi data industri menjadi big data sangat mungkin terjadi seiring perkembangan teknologi. Misalnya platform media sosial industri memiliki volume data yang sangat besar mulai dari peta-byte hingga zettabytes, yang lebih dari alam semesta fisik dalam menangkap persepsi konsumen terhadap produk mereka. Media sosial digunakan dalam bisnis sebagai alat untuk mempromosikan produk yang menawarkan rekomendasi produk kepada pengguna berdasarkan pola penelusuran dan mengumpulkan ulasan pengguna untuk memahami kebutuhan pengguna dan data di media sosial menyimpan informasi yang terkait dengan tampilan produk, suka, peringkat umpan balik, sentimen, dan emosi. Organisasi bisnis mulai memanfaatkan kumpulan data tidak terstruktur dengan bisnis terstruktur untuk mengumpulkan wawasan bisnis yang mengembangkan sejumlah besar data yang berubah menjadi Big Data. Perkembangan volume data dapat dilihat pada Tabel 8.1.

Tabel 8.1 Perkembangan Volume Data dari tahun 1930-2010 (Bhuvaneswari, 2021)

Tahun	Ukuran	Unit (bytes)	Penyimpanan
1930-1989	1 bit = 0 or 1		
	1 byte = 8 bits		Memory chips
	1 KyloBytes (KB) = 1.024 bytes	1000^1	Floppy disk
1990-2009	1 MegaByte (MB) = 1.048.576 bytes	1000^2	CD drives
	1 GigaByte (MB) = 1.073.741.824 bytes	1000^3	DVD, flash drives

Tahun	Ukuran	Unit (bytes)	Penyimpanan
2010 - sekarang	1 TeraByte (TB) = 1.099.511.627.776 bytes	1000^1	External hard disk, Cloud storage
	1 PetaByte (PB) = 1.125.899.906.842.624 bytes	1000^2	Cloud storage seperti Google drive, Drop box
	1 ExaByte (EB) = 1.000.000.000.000.000.000 bytes	1000^3	
	1 ZettaBytes (ZB) = 1.000.000.000.000.000.000.000 bytes	1000^2	
	1 YottaByte (YB) = 1.000.000.000.000.000.000.000.000 bytes	1000^3	

Selain dari volumenya data juga dilihat dari *variety* (keragaman), *velocity* (kecepatan), *veracity* (kualitas dan keakuratan) dan *value* (nilai data) dapat dilihat pada Gambar 8.1 atau biasa dikenal dengan Big Data 5V. Jenis data terbagi menjadi data terstruktur, tidak terstruktur dan semistruktur. Beberapa contoh format data dapat dilihat pada Tabel 8.2. Selanjutnya velocity mengacu pada seberapa cepat data dihasilkan dan bergerak. Perusahaan sangat membutuhkan data yang cepat untuk dianalisis dan selanjutnya proses pengambilan keputusan. Kecepatan data dihasilkan dan bergerak terbagi menjadi beberapa misalnya dalam batch, neartime, realtime sampai data stream. Data stream dimana data mengalir terus menerus dan dianalisis dalam waktu nyata. Data stream mungkin tak terbatas, setiap item memiliki stempel waktu, dan jadi urutan temporal. Terdapat dua kendala utama dalam data stream yaitu alirannya besar dan cepat, dan perlu mengekstrak informasi secara nyata. Sampai saat ini sudah banyak penelitian ke dalam data stream atau Big Data Stream Mining.



Gambar 8.1 Karakteristik Big Data

Tabel. 8.2 Contoh Data Berdasarkan Jenis Format (Bhuvaneswari, 2021)

Format Data	Domain	Sumber Data	Ukuran
Data Terstruktur	Data transaksi bisnis	Database Bank, Rumah Sakit, Smart City	KB sampai GB
Data Tidak Terstruktur	Data dari Sosial Media dapat berupa teks, audio, video dan gambar	Percakapan Whatsapp, Facebook, Twitter maupun Instagram	MB sampai PB
Data semi-struktur	Metadata dari database, data yang berasal dari mesin	Perangkat IoT – Smart Meter	KB sampai MB

Value atau nilai yang didapatkan dari big data. Nilai statistic, korelasi, kejadian, informasi-informasi yang lain yang didapat sangat tergantung dengan tujuan analisis perusahaan maupun peneliti. Informasi yang ingin didapatkan juga harus sesuai dengan pendekatan yang dilakukan. Big Data tidak hanya dilihat dari volume data namun juga menggambarkan alat dan platform yang digunakan. Big Data juga merupakan sebuah subjek lengkap dengan

alat, teknik, dan kerangka kerja. Selain itu juga bisa diterjemahkan sebagai teknologi yang berhubungan dengan kumpulan data yang besar dan kompleks yang bervariasi dalam format dan struktur data tidak sesuai dengan memori.

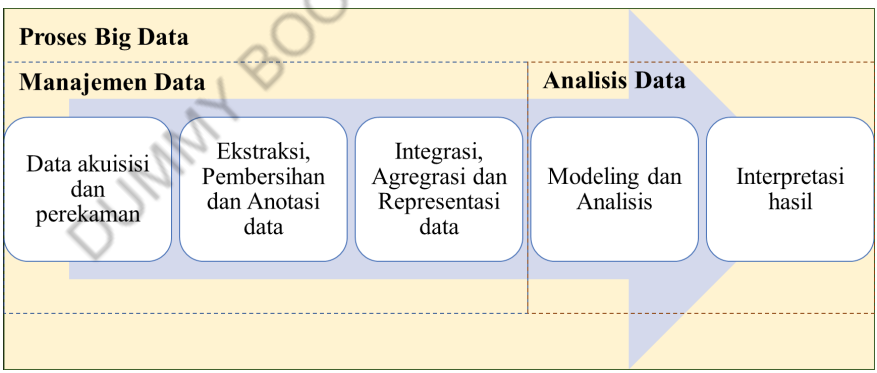
Analisis big data misalnya berasal dari industri harus melibatkan pemikiran manusia agar informasi yang ingin dicari dapat ditemukan. Analisis data berfokus pada kejadian tertentu yang ingin diamati. Metode yang digunakan juga harus disesuaikan dengan tujuan penggalian informasi yang dilakukan. Big data sangat erat dengan pendekatan statistika dan juga data mining. Metode statistika banyak digunakan dibidang industri baik dalam praktik langsung maupun penelitiannya. Banyak bidang dalam Teknik industri yang dapat menggunakan metode statistika dalam menganalisis, menyajikan data maupun menyimpulkan informasi yang ada pada data. Beberapa bidang dalam Teknik industri seperti pemasaran, sumber daya manusia, perancangan produk, sistem produksi, maupun rekayasa kualitas dapat menggunakan statistika sebagai pendekatan dalam menganalisis data.

Perusahaan besar mitra jurusan Teknik Industri Universitas Trisakti, PT Komatsu Indonesia (KI) memiliki *Quality Control Circle* (QCC) dalam menjaga setiap proses dalam perusahaan dapat berjalan sesuai standar kualitas yang diterapkan. QCC dipakai tidak hanya berfokus pada permasalahan kualitas produk namun juga permasalahan di semua bidang perusahaan. Tahapan QCC banyak memakai metode statistika dalam menganalisis dan menyajikan hasil pengamatan data yang telah dilakukan. Teknik statistika biasanya dipakai untuk data yang sedikit untuk jumlah data yang besar maka dilakukan Teknik sampling. Teknik sampling digunakan agar proses analisis data tidak membutuhkan biaya yang besar namun juga masih memiliki banyak keterbatasan seperti bias dan ragam. Seiring dengan peningkatan jumlah data seiring dengan berkembangnya perusahaan maka bukan suatu yang mustahil untuk metode analisis big data yang lain dipakai di PT KI.

Pendekatan data mining juga dapat digunakan lebih lanjut dalam mengidentifikasi pola informasi yang tersembunyi dalam data perusahaan misalnya di PT KI. Data mining dapat menangani data yang besar (big data) sehingga memungkinkan untuk tidak dilakukan sampling seperti pendekatan statistika. Metode dalam data mining sangat banyak yang dapat digunakan dalam praktik perusahaan maupun dalam bidang penelitian. Contohnya seperti

pada PT KI, metode klasifikasi dan asosiasi dapat digunakan untuk identifikasi kelas kualitas produk berdasarkan kecacatan yang ditemukan. Selain itu juga dapat digunakan sebagai klasifikasi kinerja sumber daya manusia dalam penentuan insentif maupun gaji dan masih banyak digunakan di semua bidang. Banyak kelebihan data mining namun juga memiliki keterbatasan yaitu hanya dapat digunakan pada data terstruktur. Data yang tidak terstruktur memungkinkan dianalisis namun harus melalui tahap data preprosessing agar menjadi data terstruktur dan siap diolah. Sedangkan data biasanya di perusahaan kebanyakan merupakan data tidak terstruktur.

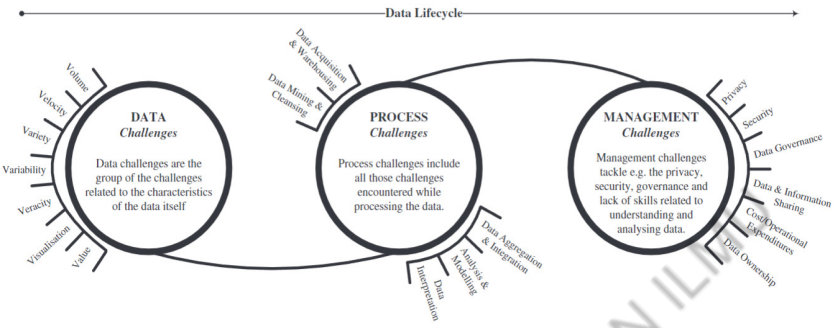
Big data tidak berharga apabila nilai potensialnya tidak dimanfaatkan untuk mendorong pengambilan keputusan. Sebelum digunakan dalam pengambilan keputusan maka harus dilakukan proses analisis big data terdahulu untuk menggali informasi yang ada didalamnya. Secara umum proses mendapatkan informasi dalam big data dapat dilihat pada Gambar 8.2. Lima tahap proses analisis big data dibagi menjadi dua sub proses utama yaitu manajemen data dan analisis data. Data manajemen melibatkan proses dan teknologi pendukung untuk memperoleh dan menyimpan data serta menyiapkan dan mengambilnya untuk analisis. Analisis data mengacu pada teknik yang digunakan untuk menganalisis dan memperoleh informasi dari data besar.



Gambar 8.2 Proses Big Data (Gandomi & Haider, 2015)

Meskipun manfaat big data sangat baik dalam pendukung keputusan perusahaan namun tetap ada banyak tantangan yang harus diatasi untuk sepenuhnya mewujudkan potensi darinya. Beberapa tantangan ini berasal karakteristik big data, metode

maupun model analisis yang ada, dan juga beberapa melalui keterbatasan sistem pemrosesan data saat ini. Tantangan tersebar baik dalam big data itu sendiri, prosesnya maupun manajemen data dapat dilihat pada Gambar 8.3.

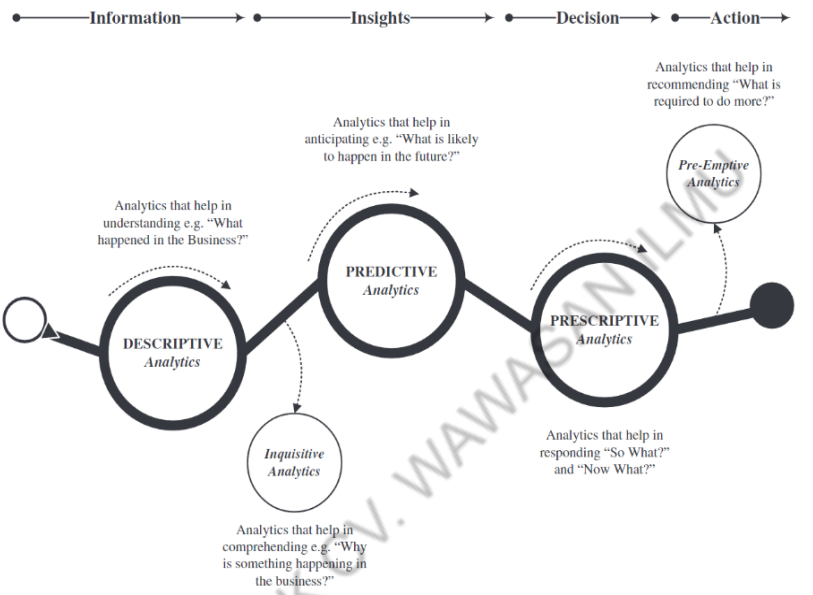


Gambar 8.3 Tantangan Big Data (Sivarajah et al., 2017)

Berbagai metode dalam analisis big data dapat dilihat pada Gambar 8.4. Pemilihan metode dapat disesuaikan dengan tujuan keputusan yang ingin diambil oleh perusahaan maupun peneliti. Beberapa metode tersebut antara lain:

1. Analisis deskriptif, menganalisis data guna mendapatkan informasi untuk menentukan keadaan situasi bisnis perusahaan, misalnya melihat perkembangan, pola dan pengecualian, maupun kejadian luar biasa yang terjadi guna membuat tindakan preventif kerugian dan merumuskan strategi peningkatan.
2. Analisis ingin tahu (*inquisitive*) adalah tentang menyelidiki data untuk mengesahkan/menolak bisnis proposisi, misalnya, penelusuran analitis ke dalam data, statistic analisis, analisis factor. Analisis ini lebih berfokus pada mengetahui penyebab terjadinya suatu kejadian.
3. Analitik prediktif berkaitan dengan peramalan dan model statistic untuk menentukan kemungkinan masa depan. Model prediksi dapat membantu perusahaan misalnya dalam melihat permintaan di masa mendatang sehingga dapat menyiapkan bahan baku maupun sumber daya yang lain agar permintaan dapat terpenuhi.
4. Analisis preskriptif adalah tentang pengoptimalan dan pengujian acak untuk menilai bagaimana bisnis meningkatkan tingkat layanan mereka sambil menurun biaya.

5. Analisis pre-emptive adalah tentang memiliki kapasitas untuk mengambil tindakan pencegahan pada peristiwa yang mungkin tidak diinginkan mempengaruhi organisasi kinerja nasional, misalnya, mengidentifikasi kemungkinan bahaya dan merekomendasikan strategi mitigasi jauh ke depan.



Gambar 8.4 Metode dalam Analisis Big Data (Sivarajah et al., 2017)

Big Data analytics transformation

Dalam menilai alasan mengapa beberapa organisasi condong ke analitik Big Data, pemahaman konkret analisis tradisional diperlukan. Analitis tradisional metode termasuk kumpulan data terstruktur yang secara berkala dinyatakan untuk tujuan tertentu. Model data yang umum digunakan untuk mengelola dan memproses aplikasi komersial adalah model relasional. Relational Database Management Systems (RDBMS) menyediakan "*user-friendly query languages*" dan memberikan kesederhanaan bahwa jaringan atau hierarki lainnya model tidak dapat mengantarkan. Dalam sistem ini adalah tabel, masing-masing dengan nama unik, di mana data terkait disimpan dalam baris dan kolom.

Aliran data ini diperoleh melalui integrasi database dan memberikan pengaruh pada penggunaan yang dimaksudkan dari

kumpulan data. Mereka tidak memberikan banyak keuntungan untuk tujuan tersebut menciptakan produk atau layanan baru seperti yang dilakukan Big Data mengarah pada transofrmasi analitik Big Data.

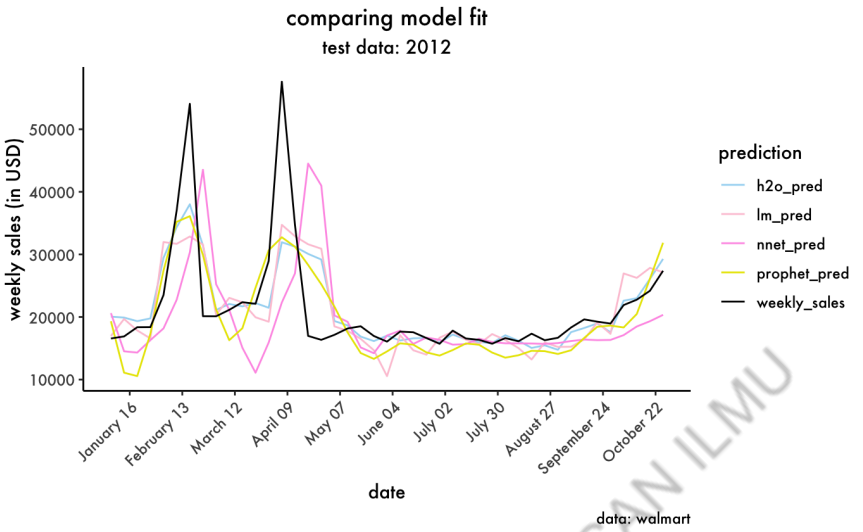
Penggunaan perangkat seluler yang sering, Web 2.0 dan pertumbuhan di Internet of Things adalah beberapa di antara alasan dibalik organisasi yang ingin mengubah analitis proses. Organisasi tertarik pada analitik data besar karena itu menyediakan sarana untuk mendaptkan data waktu nyata, cocok untuk meningkatkan operasi bisnis.

8.4 Studi Kasus Big Data

Walmart

Bagaimana Big Data Digunakan untuk Mendorong Kinerja Supermarket (Marr, 2016)

Walmart adalah perusahaan retail terbesar di dunia berdasarkan pendapatan, dengan lebih dari dua juta karyawan dan 20.000 toko di 28 negara. Dengan operasi pada skala ini, tidak mengherankan bahwa mereka telah lama melihat nilai dalam analisis data. Pada tahun 2004, ketika Badai Sandy menghantam AS, mereka menemukan bahwa wawasan tak terduga dapat terungkap ketika data dipelajari secara keseluruhan, bukan sebagai kumpulan individu yang terisolasi. Mencoba memperkirakan permintaan pasokan darurat secara langsung dari Badai Sandy yang mendekat, CIO Linda Dillman muncul dengan beberapa hasil statistik yang mengejutkan. Selain senter dan peralatan darurat, cuaca buruk yang diperkirakan telah menyebabkan peningkatan penjualan stroberi Pop Tart di beberapa lokasi lain. Persediaan tambahan ini adalah dikirim ke toko-toko di jalur Hurricane Frances pada tahun 2012, dan dijual sangat baik. Walmart telah mengembangkan departemen Big Data dan analitik mereka secara signifikan sejak itu, terus menjadi yang terdepan. Pada tahun 2015, perusahaan mengumumkan bahwa mereka sedang dalam proses menciptakan cloud data pribadi terbesar di dunia, untuk memungkinkan pemrosesan 2.5 petabyte informasi setiap jam. Contoh prediksi penjualan barang Walmart dapat dilihat pada Gambar 8.5.



Gambar 8.5 Prediksi Sales Walmart (Heschmeyer, 2019)

8.5 Hadoop

Apache Hadoop adalah kumpulan utilitas perangkat lunak sumber terbuka yang memfasilitasi penggunaan jaringan banyak komputer untuk memecahkan masalah yang melibatkan sejumlah besar data dan komputasi. Ini menyediakan kerangka kerja perangkat lunak untuk penyimpanan terdistribusi dan pemrosesan data besar menggunakan model pemrograman MapReduce. Hadoop awalnya dirancang untuk cluster komputer yang dibangun dari perangkat keras komoditas, yang masih digunakan secara umum. Sejak itu juga ditemukan digunakan pada kelompok perangkat keras kelas atas. Semua modul di Hadoop dirancang dengan asumsi mendasar bahwa kegagalan perangkat keras adalah kejadian umum dan harus ditangani secara otomatis oleh kerangka kerja.

8.6 Map Reduce

MapReduce adalah model pemrograman rilisan Google yang ditujukan untuk memproses data berukuran raksasa secara terdistribusi dan paralel dalam cluster yang terdiri atas ribuan komputer. Dalam memproses data, secara garis besar MapReduce dapat dibagi dalam dua proses yaitu proses Map dan proses Reduce.

Kedua jenis proses ini didistribusikan atau dibagi-bagikan ke setiap komputer dalam suatu cluster (kelompok komputer yang salih terhubung) dan berjalan secara paralel tanpa saling bergantung satu dengan yang lainnya. Proses Map bertugas untuk mengumpulkan informasi dari potongan-potongan data yang terdistribusi dalam tiap komputer dalam cluster. Hasilnya diserahkan kepada proses Reduce untuk diproses lebih lanjut. Hasil proses Reduce merupakan hasil akhir yang dikirim ke pengguna.

Kemajuan teknologi dalam beberapa akhir dekade telah menyebabkan ledakan ukuran data, meskipun sekarang ada lebih banyak untuk bekerja dengan kecepatan di mana ini volume data tumbuh melebihi sumber daya komputasi tersedia. Map Reduce, paradigma pemrograman yang digunakan untuk memproses kumpulan data besar di linkgungan terdistribusi adalah dipandang sebagai pendekatan untuk menangani peningkatan permintaan untuk sumber daya komputasi.

Ada dua fungsi mendasar dalam paradigma, peta dan fungsi Reduce. Fungsi Peta dijalankan menyortir dan memfilter, secara efektif mengubah kumpulan data menjadi yang lain & fungsi Reduce mengambil output dari Map berfungsi sebagai input, lalu menyelesaikan pengelompokan dan agregasi operasi untuk menggabungkan set data tersebut menjadi set yang lebih kecil.

Sumber : https://www.teknologi-bigdata.com/2013/02/mapreduce-besar-dan-powerful-tapi-tidak.html

8.7 Latihan Soal

1. Jelaskan resiko dan keuntungan dari big data dan berikan contoh!
2. Jelaskan apa keuntungan menggunakan Hadoop dan berikan contoh!
3. Jelaskan apa keuntungan menggunakan Map Reduce dan berikan contoh!

BAB IX
DATA WAREHOUSE DAN OLAP

9.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk mengenalkan prinsip data warehouse, OLAP (*Online analitical processing*), *Extract, Transform* dan *Load* (ETL))

Materi yang diberikan :

1. Data Warehouse
2. OLAP (*Online analytical processing*)
3. *Extract, Transform* dan *Load* (ETL)

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai prinsip data warehouse, OLAP *(Online analytical processing*), *Extract, Transform* dan *Load* (ETL) dan Intelijensia Bisnis *(Business Intelligence)*. Capaian pembelajaran lulusan adalah mahasiswa mampu mengimplementasikan prinsip data warehouse, OLAP (*Online analytical processing*), *Extract, Transform* dan *Load* (ETL) pada sistem.

9.2. Data Warehouse

Data warehouse adalah koleksi dari data yang *subject-oriented*, terintegrasi, *time- variant*, dan *nonvolatile*, dalam mendukung proses pembuatan keputusan. Sering diintegrasikan dengan berbagai sistem aplikasi untuk mendukung pemrosesan informasi dan analisis data dengan menyediakan platform untuk *historical data*. *Data warehousing*: proses konstruksi dan penggunaan *data warehouse*.

Data warehouse bersifat subject oriented. Data warehouse diorganisasikan diseputar subjek-subjek utama seperti customer, produk, sales. Fokus pada pemodelan dan analisis data untuk pembuatan keputusan, bukan pada operasi harian atau pemrosesan

transaksi. Menyediakan sebuah tinjauan sederhana dan ringkas seputar subjek tertentu dengan tidak mengikut sertakan data yang tidak berguna dalam proses pembuatan keputusan.

Data warehouse bersifat terintegrasi. Dikonstruksi dengan mengintegrasikan banyak sumber data yang heterogen, relational database, flat file, on-line transaction record. Teknik *data cleaning* dan *data integration* digunakan untuk menjamin konsistensi dalam konvensi-konvensi penamaan, struktur pengkodean, ukuran-ukuran atribut dll diantara sumber data yang berbeda. Contoh: Hotel price: currency, tax, breakfast covered, dll. Data dikonversi ketika dipindahkan ke *warehouse*.

Data warehouse bersifat *time variant*. Data disimpan untuk menyediakan informasi dari perspektif *historical*, contoh 5-10 tahun yang lalu. Struktur kunci dalam *data warehouse* mengandung sebuah elemen waktu, baik secara ekspisit atau secara implisit. Tetapi kunci dari data operasional bisa mengandung elemen waktu atau tidak.

Data warehouse bersifat *non volatile*. Data warehouse adalah penyimpanan data yang terpisah secara fisik yang ditransformasikan dari lingkungan operasional. Data warehouse tidak memerlukan pemrosesan transaksi, recovery dan mekanisme kontrol konkurensi. Biasanya hanya memerlukan dua operasi dalam pengaksesan data, yaitu initial loading of data dan access of data.

9.3 OLAP (*Online analitical processing*)

OLAP (*Online analitical processing*) adalah operasi basis data untuk mendapatkan data dalam bentuk kesimpulan dengan menggunakan agregasi sebagai mekanisme utama. Ada 3 tipe OLAP yaitu *Relational* OLAP (ROLAP), *Multidimensional* OLAP (MOLAP), *Hybrid* OLAP (HOLAP).

Datawarehouse dibandingkan dengan Operasional DBMS. OLTP (on-line transaction processing) bertugas untuk relasional tradisional Data Base Management System, operasi harian, pembelian, persediaan, banking, manufacturing, payroll, registration, accounting, dan sebagainya. OLAP (on-line analytical processing) merupakan tugas utama dari sistem data warehouse, untuk menganalisa data dan membuat keputusan.

Perbedaan antara OLTP dan OLAP adalah :

Tabel 9.1 Perbedaan OLTP dan OLAP

Kriteria	OLTP	OLAP
Orientasi Pengguna dan Sistem	Berorientasi terhadap pelanggan	Berorientasi terhadap pasar
Konten Data	current, detailed	historical, consolidated
Disain Database	ER + application	star + subject
View	current, local	vs. evolutionary, integrated
Access patterns	update	read-only but complex queries

Star schema : Sebuah tabel fakta di tengah-tengah dihubungkan dengan sekumpulan tabel-tabel dimensi.

Snowflake schema : perbaikan dari skema star ketika hirarki dimensional dinormalisasi ke dalam sekumpulan tabel-tabel dimensi yang lebih kecil

Fact constellations : Beberapa tabel fakta dihubungkan ke tabel-tabel dimensi yang sama, dipandang sebagai kumpulan dari skema star, sehingga dinamakan skema galaksi atau *fact constellation*.

OLAP adalah teknologi untuk memproses analisa informasi berdasarkan datawarehouse yang memudahkan pengguna untuk mengobservasi data dari multidimensional melalui operasi *slice*, *dice*, *rollup/drill down* dan *rotate*. Slice adalah untuk memilih kelompok dari satu dimensi dari multidimensional array, dice untuk memilih beberapa area dimensi anggota dari satu dimensi, roll up mengambil low layer detail data untuk menjumlahkan kumpulan high layer. Oleh karena itu kombinasi dari datawarehouse dan OLAP dapat secara efektif memecahkan persoalan bagaimana untuk menangani data yang banyak dalam pengambilan keputusan (Liangzhoong, 2012).

Datawarehouse adalah kumpulan dari data yang diekstrak dari beberapa sistem operasi yang dilakukan transformasi dan *load* data yang konsisten untuk dianalisis. *Data Mart* adalah bagian penting dari *data warehouse*. Pembuatan *Data warehouse* dan *Data Mart* masuk ke dalam tahap perencanaan dalam Sistem Intelijensia Bisnis.

Arsitektur *Data Mart* Sistem Intelijensia Bisnis untuk Agroindustri Susu Skala Menengah menggunakan *star schema.* Pada *star schema*, seluruh informasi untuk suatu hirarki diletakkan di tabel yang sama. *Data warehouse* dan *Data Mart* dibuat dengan menggunakan SQL *Server* 2008 *Visual Studio*.

Data Mart adalah bagian dari data yang penting dari *data warehouse* pusat. *Data Mart* dapat digunakan untuk memberikan informasi ke *data warehouse* pusat. Ketika *data warehouse* didesain untuk melayani kebutuhan perusahaan, *data mart* juga melayani kebutuhan dari bisnis unit khusus, fungsi, proses atau aplikasi. Karena *data mart* berhubungan langsung kebutuhan bisnis yang khusus, beberapa bisnis dapat melewatkan *data warehouse* dan membuat *data mart*. *Data Mart* adalah *repository* untuk data yang digunakan untuk Intelijensia Bisnis. *Data Mart* secara periodically menerima data dari *online transactional processing* (OLTP) system. Arsitektur *data mart* menggunakan skema *star* atau skema *snowflake*.

Struktur operasional utama dalam OLAP didasarkan pada konsep yang disebut kubus (*cube*) (Turban *et al.*, 2011). Kubus (*cube*) didalam OLAP adalah struktur data multidimensional (*actual* atau *virtual*) yang memungkinkan analisis data yang cepat. Juga dapat didefinisikan sebagai kemampuan dari memanipulasi dan menganalisis data secara efisien dari berbagai perspektif.

Customer life time value (CLV) menggambarkan nilai sekarang dari arus laba masa depan (*net present value of the stream of future profit*) yang diharapkan selama pembelian seumur hidup pelanggan (Kotler, 2009). Perusahaan harus mengurangi dari pendapatan yang diharapkan biaya untuk menarik, menjual dan melayani pelanggan itu. Menghitung CLV mempunyai banyak aplikasi dan beberapa pengarang telah mengembangkan model- model untuk aplikasinya seperti alokasi sumber daya.

9.4 Extract, Transform dan Load (ETL)

Extract, Transform dan *Load* (ETL) melakukan proses ekstrak data untuk mengcopy dari satu atau lebih sistem OLTP, menampilkan semua data yang akan dibersihkan untuk ditransform menjadi data dengan format yang konsisten dan diload data yang telah dibersihkan menjadi *data mart* (Larson,2009). Pembersihan data

menghilangkan ketidakkonsistenan dan *error* dari data transaksi untuk kepentingan konsistensi untuk penggunaan *data mart*.

Extract Transform Load (ETL) mengacu pada perangkat lunak yang ditujukan untuk tampil disebuah otomatisasi dengan cara tiga fungsi utama : ekstraksi, transformasi dan loading data ke data warehouse (Vercelis, 2009). Spoon merupakan *integrated development environment* (IDE) yang berupa *graphical user interface* (GUI) yang digunakan untuk merancang, menyunting dan menjalankan *job* dan *transformation. Online transactional processing* (OLTP) adalah teknologi untuk mengelola aplikasi yang berorientasi pada transaksi. OLAP adalah teknologi untuk menjawab kebutuhan analitik. Setiap step memiliki kemampuan untuk memberikan penjelasan mengenai baris yang dikeluarkannya. Penjelasan dari baris ini disebut dengan *row metadata. Row metadata* mengandung informasi seperti nama, tipe data, lebar data, lebar presisi desimal. (Mulyana, 2014).

9.5 Studi Kasus

Analisis *Cube* (Fitriana et. al., 2018)

Pemodelan *Customer Relationship Management* (CRM)

Nilai CLVk = (NRk x WRk) + (NFk x WFk) + (NMk x WMk)

CLVk = *Customer Lifetime Value* = Nilai Siklus Hidup Konsumen ke k

NRk = Normalisasi *Recency*

WRk = Bobot *Recency*

NFk = Normalisasi *Frequency* transaksi penjualan WFk = Bobot *Frequency*

NMk = Normalisasi hasil penjualan WMk = Bobot *Monetary*

k = pedagang roti

Metode yang digunakan untuk membangun customer relationship management untuk R*Bakery* adalah metode analisis pemasaran RFM (*Recency, Frequency, Monetary*) dan *Customer Lifetime Value* (CLV) yang digunakan untuk menentukan siklus hidup konsumen yang hasilnya dapat menentukan pedagang potensial oleh R Bakery. Berikut adalah rumus matematika untuk pemodelan CRM dengan RFM dan CLV (Rina *et al.,* 2012).

Tabel 9.2 Tabel bobot CLV

WR_k	0,3
WF_k	0,35
WM_k	0,35

Keterangan kategori RFM :
NRk = *untuk (high, medium, low*) nilai NRk adalah (1,2,3)
NFk = untuk (*high, medium, low*) nilai NFk adalah (3,2,1)
NMk = untuk (*high, medium, low*) nilai NFk adalah (3,2,1)

Tabel 9.3 Kategori RFM

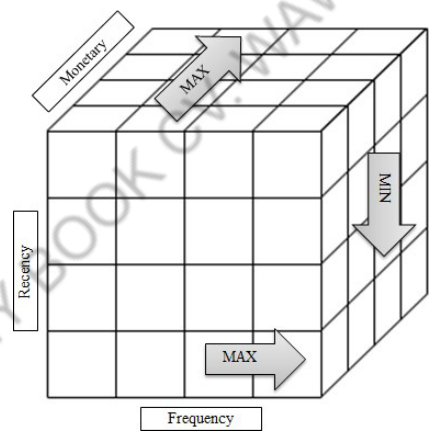
	Low	Medium	High
Recency	1	2-3	4-5
Frequency	15-20	21-25	26-30
Monetory	9.000.000-10.00000	10.100.000-11.500.000	11.600.000-12.700.000

Tabel 9.4 Hasil RFM pada bulan September 2016

Nama Pedagang	Frekuensi	Monetory	Recency	Kelas Frequency	Kelas Monetory	Kelas Recency
JJ	30 hari/bln	10.258.000	1	High	Medium	Low
MR	30 hari/bln	10.237.000	1	High	Medium	Low
AQ	24 hari/bln	11.447.500	2	Medium	Medium	Medium
BD	25 hari/bln	12.384.500	2	Medium	High	Medium
LM	24 hari/bln	12.025.500	2	Medium	High	Medium
DF	23 hari/bln	12.497.500	3	Medium	High	Medium
YY	26 hari/bln	11.968.000	3	High	High	Medium
CK	17 hari/bln	10.105.000	5	Low	Low	High
MK	25 hari/bln	10.546.500	2	Medium	Medium	Medium
IR	24 hari/bln	11.264.000	2	Medium	Medium	Medium
ST	27 hari/bln	12.041.500	3	High	High	Medium
SM	28 hari/bln	12.399.000	3	High	High	Medium
AW	30 hari/bln	10.168.000	1	High	Low	Low
UD	30 hari/bln	11.565.000	1	High	Medium	Low
FA	30 hari/bln	9.823.000	1	High	Low	Low

Nama Pedagang	Frekuensi	Monetory	Recency	Kelas Frequency	Kelas Monetory	Kelas Recency
SW	18 hari/bln	10.542.000	5	Low	Medium	High
BL	21 hari/bln	11.185.000	4	Medium	Medium	High
AT	23 hari/bln	10.749.000	3	Medium	Medium	Medium
PN	20 hari/bln	10.177.500	4	Low	Low	High
MY	19 hari/bln	12.097.000	4	Low	High	High
AJ	30 hari/bln	12.625.500	1	High	High	Low
AD	29 hari/bln	11.250.000	2	High	Medium	Medium
DL	22 hari/bln	11.610.000	4	Medium	High	High
KW	30 hari/bln	9.677.500	1	High	Low	Low

Dari hasil RFM tersebut pada Tabel 7 diaggregasi sebagai logical *cube* yang berbasis OLAP. Untuk meminimasi *Recency*, maksimasi *Frequency*, dan maksimasi *monetary* dikumpulkan dengan CLV (*Customer Lifetime Value*) terbaik dari bulan September 2016. Logical *cube* RFM disajikan pada Gambar 17 dan ringkasan dari hasil CLV disajikan pada Tabel 8.



Gambar 9.1 Cube RFM

Usulan Perbaikan CLV (*Customer Lifetime Value*)

CLV menghasilkan *output* berupa *ranking* dari pedagang di R *Bakery*. Usulan yang diberikan adalah dengan memberikan penghargaan kepada pedagang *ranking* 1 hingga 3 dan memberikan semangat dan motivasi kepada pedagang yang *ranking* dibawah 3. Usulan ini diberikan agar pada pedagang lebih semangat dalam memasarkan produk roti R Bakery.

Tabel 9.5 Tabel ranking CLV bulan September 2016

Nama Pedagang	Frekuensi	Monetary	Recency	Kelas Frekuensi	Kelas Monetory	Kelas Recency	CLV	Ranking
MY	19 hari/bln	12.097.000	4	Low	High	High	3	1
JJ	30 hari/bln	10.258.000	1	High	Medium	Low	2,65	2
MR	30 hari/bln	10.237.000	1	High	Medium	Low	2,65	2
UD	30 hari/bln	11.565.000	1	High	Medium	Low	2,65	2
SW	18 hari/bln	10.542.500	5	Low	Medium	High	2,65	2
BD	25 hari/bln	12.384.500	2	Medium	High	Medium	2,35	3
LM	24 hari/bln	12.025.500	2	Medium	High	Medium	2,35	3
DF	23 hari/bln	12.497.500	3	Medium	High	Medium	2,35	3
BL	21 hari/bln	11.185.000	4	Medium	Medium	Medium	2,35	3
CK	17 hari/bln	10.105.000	5	Low	Low	High	2,3	4
PN	20 hari/bln	10.177.500	4	Low	Low	High	2,3	4
YY	26 hari/bln	11.968.000	3	High	High	Medium	2,05	5
ST	27 hari/bln	12.041.500	3	High	Low	Low	2,05	5
SM	28 hari/bln	12.399.000	3	High	Medium	Low	2,05	5
AQ	24 hari/bln	11.447.500	2	Medium	Medium	Medium	2	6
MK	25 hari/bln	10.546.500	2	Medium	Medium	Medium	2	6
IR	24 hari/bln	11.264.000	2	Medium	Medium	Medium	2	6

Nama Pedagang	Frekuensi	Monetary	Recency	Kelas Frekuensi	Kelas Monetory	Kelas Recency	CLV	Ranking
AT	23 hari/bln	10.749.000	3	Medium	Medium	Medium	2	6
AD	29 hari/bln	11.250.000	2	High	Medium	Medium	1,7	7
AJ	30 hari/bln	12.625.500	1	High	High	Low	1,7	7
DL	22 hari/bln	11.610.000	4	Medium	High	High	1,3	8
AW	30 hari/bln	10.168.000	2	High	Low	Low	1	9
FA	30 hari/bln	9.823.000	1	High	Low	Low	1	9
KW	30 hari/bln	9.677.500	1	High	Low	Low	1	9

Model *cube* dan *data mining* yang menampilkan hasil visual yang akan membantu dalam pengambilan keputusan untuk meningkatkan pemasaran. Hasil *data mining clustering k-means* adalah terdiri dari 83% *cluster* 1 dan 17% *cluster* 2. *Cluster* 1 merupakan kategori sisa roti yang rendah dan *cluster* 2 merupakan kategori sisa roti yang tinggi. Hasil yang diperoleh dari *On Line Analitical Process Cube* adalah berupa *data warehouse* penjualan R Bakery. Dari perhitungan CLV untuk pemodelan CRM diketahui bahwa urutan pedagang dengan *ranking* CLV tertinggi adalah pedagang MY yaitu pedagang yang memiliki kelas *frequency low* namun kelas *monetary* dan kelas *recency high*.

9.6 Contoh Soal

1. Jelaskan penerapan Data Warehouse di industri manufaktur!
2. Jelaskan penerapan OLAP di industri jasa!

BAB X
SPATIAL DATA MINING

10.1Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk

- Mengenalkan tentang spatial data mining
- Mengenalkan spatial data mining framework
- Metode dalam spatial data mining

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai pemahaman mengenai spatial data mining dan aplikasinya.

10.2 Pendahuluan

Data spasial memainkan peran penting dalam pengembangan keberlanjutan sumber daya alam dan masyarakat, tidak hanya untuk memenuhi tuntutan manusia untuk mempelajari sumber daya dan lingkungan. Sumber informasi geografis juga dapat dimanfaatkan dalam semua bidang termasuk teknik industri. Informasi dalam data spasial jauh berbeda dengan data transaksional di mana data spasial lebih kaya akan informasi yang perlu didalami dan sangat luas, selain itu juga jenisnya adalah lebih bervariasi, dan strukturnya lebih kompleks; pertumbuhan fitur data spasial yang pesat dan cepat meningkatkan volumenya. Metode tradisional untuk memproses data spasial relatif tertinggal jika dibandingkan dengan pertumbuhan big data. Besar jumlah data spasial tidak dapat diinterpretasikan dan diintrepretasikan secara tepat waktu sehingga diperlukan spasial data mining untuk hal ini.

Spatial Data Mining (SDM) oleh Li Deren (Li dan Cheng 1994) merupakan Penemuan Pengetahuan dari Pendekatan basis data

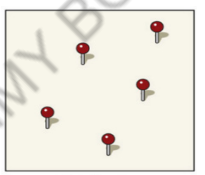
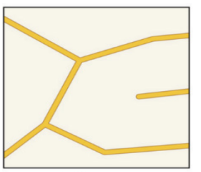
Sistem Informasi Geografis / *Knowledge Discovery from Geographical Information Systems databases* (KDG), yang terus membantu SDM dalam pengembangan lebih lanjut dari penambangan data dan geomatika. Sektor industri juga tidak lepas dari kebutuhan data spasial guna untuk berbagai kebutuhan seperti pengambilan kebijakan. Bab ini akan membahas tentang *Spatial Data Mining* dalam contoh kasus dunia industri.

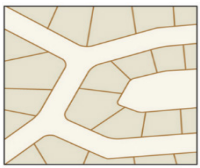
10.3 Spatial Data Mining

A. Data Spasial

Secara umum, beberapa jenis aturan geometris dapat ditemukan dari database Sistem Informasi Geografis atau *Geographic Information System* (GIS) seperti bentuk geometris, distribusi, evolusi, dan bias. Pengetahuan geometri mengacu pada fitur geometris umum dari sekelompok objek target (misalnya, volume, ukuran, bentuk). Secara umum dalam SDM, data yang dapat digunakan dibagi menjadi dua yaitu data spasial dan data nonspasial/atribut. Data spasial merupakan data yang bergeorefernsi. Struktur data spasial dapat berupa data vektor dan data raster. Data vektor dibagi menjadi tiga kategori yaitu:

Tabel 10.1 Kategori data Vektor

Kategori	Representasi	Contoh
Titik	 (Ithape, 2022)	pohon independen, pemukiman pada peta skala kecil, sekolah, kantor pemerintahan, industri
Garis	 (Ithape, 2022)	sungai, jalan nasional, jalan provinsi

Poligon	 (Ithape, 2022)	Pemukiman penduduk, danau, kota, kebun, sawah, kawasan industri

Penggunaan data vector memiliki keuntungan dan kekurangan diantaranya:

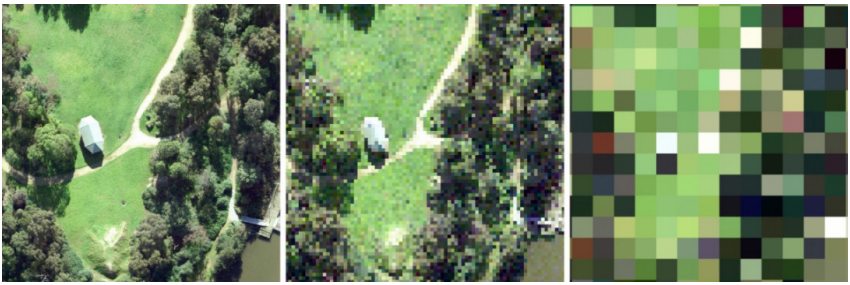
Keuntungan dari model data vektor:

a. Penyajian data cukup baik
b. Struktur datanya sangat kompak
c. Grafik data cukup akurat
d. Pemanggilan kembali, updating dan generalisasi grafik data maupun atribut dapat dilakukan dengan mudah
e. Topologi data dapat sepenuhnya dijelaskan oleh koneksi jaringan

Kelemahan model data vektor:

a. Struktur data yang cukup kompleks
b. Menggabungkan beberapa poligon dalam overlay sangat sulit
c. Menampilkan dan menyajikan data memiliki biaya yang signifikan
d. teknologinya sangat mahal, terutama untuk perangkat keras dan perangkat lunak dapat diandalkan
e. Memfilter menurut poligon tidak dimungkinkan

Sedangkan data raster merupakan data yang dapat berupa peta dan foto udara hasil penyiaman (scanning), citra penginderaan jauh dari satelit, orthophotography digital, digital elevation models (data tinggi permukaan bumi dalam format matrik ketinggian). Contoh data raster dapat dilihat pada Gambar.



Gambar 10.1 Resolusi berbeda (ukuran sel) dari gambar raster yang sama. Kiri – ukuran sel 6cm, Tengah- ukuran sel 60cm, Kanan- ukuran sel 6m (Romeijn, 2022)

Data raster memiliki keuntungan dan kelemahan dalam penggunaannya sebagai berikut:

Keuntungan dari model data raster:

a. Struktur datanya sangat sederhana
b. Mengoverlay dan menggabungkan data dan gambar sangat mudah
c. Beberapa analisis spasial dapat dilakukan dengan cukup mudah
d. Simulasi sangat mudah dilakukan karena semua data sudah dalam bentuk dan ukuran yang sama
e. Teknologi tidak mahal

Kelemahan model data raster:

a. Volume grafik data cukup besar
b. Jika kotak sel besar digunakan untuk mengurangi volume
c. Data menghilangkan beberapa informasi penting
d. Koneksi jaringan sulit, dengan kata lain, transformasi proyektif membutuhkan waktu
e. Tampilan data tidak cair dan dianggap tidak sesuai

Selanjutnya data non spasial/non graphic/atribut merupakan data penunjang yang memberikan penjelasan karakteristik dari entitas geografik yang diperlukan. Data non spasial dapat terdiri atas data kualitatif dan kuantitatif. Contoh data kualitatif misalnya nama dan alamat pemilik lahan. Sedangkan data kuantitatif dapat berupa Panjang jalan, luas daerah, harga lahan, produktivitas lahan, kualitas tanah maupun masinh banyak lainnya. Dalam bidang Teknik industri data kualitatif seperti minat beli konsumen di retail

yang dimiliki, produktivitas supplier maupun kinerja distributor sampai ke kualitas bahan baku juga dapat digunakan sebagai data penunjang.

B. Memahami Spatial Data Mining

SDM adalah proses berulang dari pengambilan keputusan yang mengacu pada data spasial. Bisa jadi dipahami paling baik dari lima perspektif berikut (Li et al., 2015):

a. Kompleks, tidak pasti, dan memiliki variasi ketika menggali informasi dari data. SDM dihasilkan dari perkembangan teknologi seperti: teknologi akses data spasial, teknologi basis data spasial, statistik spasial, dan sistem informasi spasial. Teori dan tekniknya terkait dengan penambangan data, penemuan pengetahuan, sistem basis data, analisis data, pembelajaran mesin, pengenalan pola, ilmu kognitif, kecerdasan buatan, statistik matematika, jaringan teknologi, rekayasa perangkat lunak, dll.

b. Analysis, SDM menemukan aturan yang tidak diketahui dan berguna dari jumlah besar data melalui serangkaian manipulasi interaktif, berulang, asosiatif, dan berorientasi data. Ini terutama menggunakan metode dan teknik tertentu untuk mengekstrak berbagai pola dari dataset spasial. Pola yang ditemukan menggambarkan yang ada aturan atau memprediksi tren yang berkembang, bersama dengan kepastian atau kredibilitas untuk mengukur kepercayaan, dukungan, dan minat kesimpulan yang diperoleh dari analisis. Mereka dapat membantu pengguna memanfaatkan sepenuhnya repositori spasial di bawah payung berbagai aplikasi, menekankan implementasi yang efisien dan tepat waktu respon terhadap perintah pengguna.

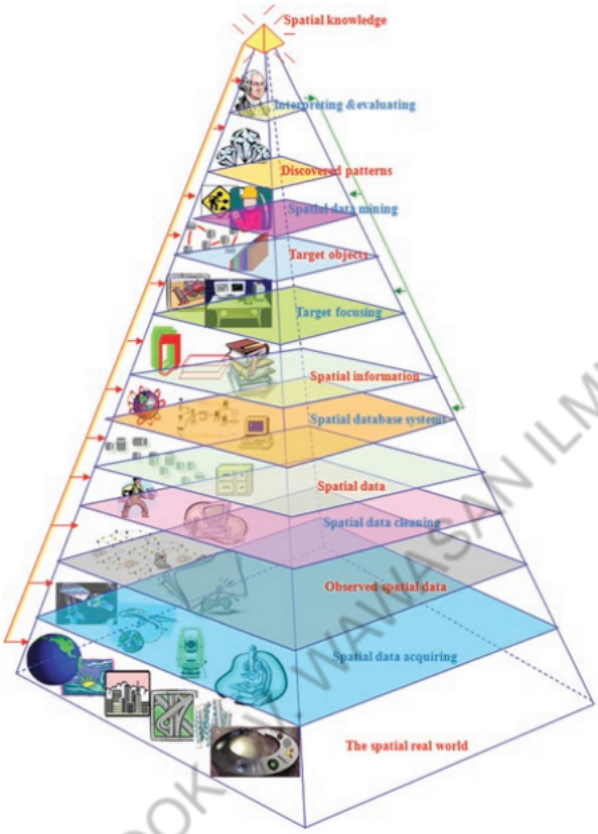
c. Logic, SDM adalah teknik lanjutan dari penalaran spasial deduktif. SDm focus pada penemuan informasi bukan pemeriksaan, dalam konteks data mining. Sebagai bagian dari inferensi deduktif, SDM merupakan alat khusus untuk penalaran spasial yang memungkinkan pengguna untuk mengawasi atau fokus pada penemuan aturan yang menarik. Alasannya bisa otomatis atau semi otomatis. Induksi digunakan untuk menemukan pengetahuan, sedangkan deduksi digunakan untuk mengevaluasi pengetahuan yang ditemukan. Algoritma penambangan merupakan kombinasi dari induksi dan deduksi.

d. Actual object operation. Model data dapat berupa hierarki, jaringan, relasional, berorientasi objek, terkait objek, semi terstruktur, atau tidak terstruktur. Format data dapat berupa data spasial vektor, raster, atau raster vektor. Data repositori adalah sistem file, database, pasar data, gudang data, dll.

   Konten data mungkin melibatkan lokasi, grafik, gambar, teks, aliran video, atau kumpulan data lainnya yang diorganisasikan bersama, seperti data multimedia dan data jaringan.

e. Source. SDM diimplementasikan pada data asli dalam database, data yang dibersihkan di gudang data, perintah terperinci, informasi dari pengguna, dan pengetahuan latar belakang dari bidang yang berlaku, yang mencakup data mentah yang disediakan oleh database spasial dan database atribut yang sesuai atau data yang dimanipulasi yang disimpan di gudang data spasial, instruksi lanjutan dikirim oleh pengguna ke pengontrol, dan berbagai pengetahuan ahli yang berbeda bidang yang disimpan dalam basis pengetahuan. Dengan bantuan jaringan, SDM dapat memecah batasan lokal data spasial, tidak hanya menggunakan data spasial di sektornya sendiri, tetapi cakupan yang lebih besar atau bahkan semua data di bidang luar angkasa dan bidang terkait. Ini juga dapat tersedia untuk menemukan ruang yang lebih universal pengetahuan dan menerapkan spasial online data mining (SOLAM). Bertemu kebutuhan pengambilan keputusan, SDM memanfaatkan desentralisasi heterogeny sumber data, dengan informasi dan pengetahuan yang diambil secara tepat waktu dan akurat melalui analisis data menggunakan kueri dan alat analisis pelaporan modul.

Seperti konsep dasar dalam data mining maka proses SDM juga sama dimulai dari penyiapan data yang akan digali informasinya. Penyiapan data (*data preparation*) bertujuan untuk memahami pengetahuan sebelumnya di bidang aplikasi, menghasilkan target dataset, pembersihan data, dan penyederhanaan data. Setelah data siap maka dilakukan data mining dimana pemilihan fungsi dan algoritma data mining; mencari pengetahuan yang diminati dalam bentuk aturan dan pengecualian tertentu: asosiasi spasial, karakteristik, klasifikasi, regresi, pengelompokan, urutan, prediksi, dan dependensi fungsi), dan pasca-pemrosesan data mining (interpretasi, evaluasi, dan aplikasi). Proses dalam SDM dapat digambarkan dengan piramida SDM seperti Gambar.



Gambar 10.2 Piramida SDM

C. Penggunaan Spatial Data Mining

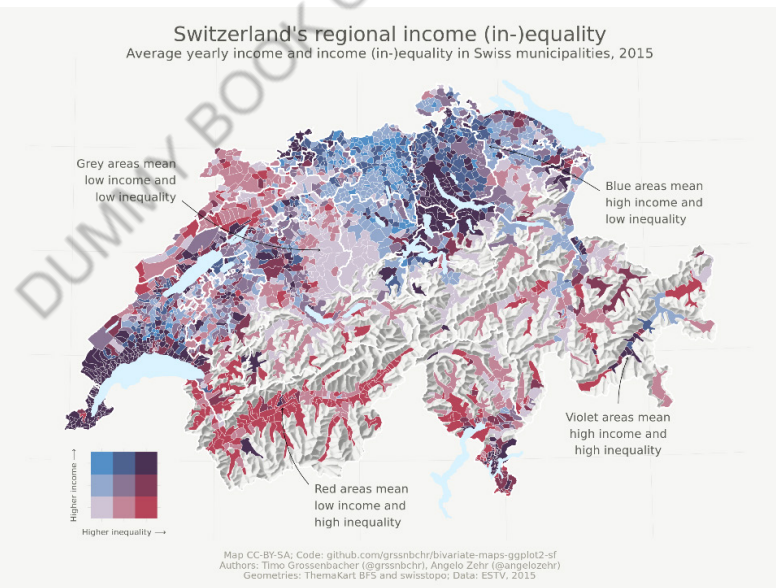
Berbagai pengetahuan dapat ditemukan dari dataset spasial. Berbagai pengetahuan dapat digali dengan menggunakan algoritma yang terkait seperti asosiasi, karakteristik, diskriminasi, pengelompokan, klasifikasi, serial, prediksi, dan ketergantungan fungsional. Lebih lengkap terkait contoh penggunaan algoritma-algoritma dalam penggalian informasi pada data spasial dapat dilihat pada Tabel. Algoritma yang digunakan dapat disesuaikan dengan jenis pengetahuan juga tergantung pada jenis tugas ataupun *spatial data mining*. Berbagai algoritma juga dapat dikombinasikan saat menyelesaikan permasalahan yang kompleks.

Tabel 10.2 Pengetahuan Spasial yang Dapat Digali (Li et al., 2016)

Pengetahuan	Interpretasi	Contoh
Association rule	Asosiasi logika di antara kumpulan entitas yang berbeda yang mengasosiasikan satu atau lebih objek dengan objek lain untuk mempelajari frekuensi item terjadi bersama-sama dalam database.	Hujan (lokasi, jumlah hujan) => Longsor (lokasi, kejadian), dengan nilai support 76%, confidence 98%, interest 51%
Characteristics rule	Sebuah fitur umum dari jenis entitas, atau beberapa jenis entitas, untuk meringkas fitur serupa dari objek di kelas target.	Karakterisasi objek tanah yang serupa dalam satu set besar penginderaan jauh gambar-gambar.
Discriminate rule	Fitur berbeda yang membedakan satu entitas dari yang lain entitas dan untuk membandingkan inti fitur objek antara target kelas dan kelas yang kontras.	Bandingkan harga tanah di pinggiran kota daerah dengan harga tanah di pusat kota.
Clustering rule	Aturan segmentasi yang mengelompokkan satu set objek berdasarkan mereka kesamaan tanpa pengetahuan apapun tentang apa yang menyebabkan pengelompokan dan berapa banyak kelompok yang ada. Sebuah cluster adalah kumpulan objek data yang mirip satu sama lain dalam cluster yang sama dan berbeda objek di cluster lain.	Kelompokkan lokasi kejahatan untuk menemukan distribusi pola.
Classification rule	Aturan yang menentukan apakah suatu entitas milik kelas tertentu dengan jenis dan variabel yang ditentukan, atau satu set kelas.	Klasifikasi gambar penginderaan jauh berdasarkan spektrum dan data GIS.
Serial rule	Aturan dibatasi temporal bahwa berhubungan entitas atau fungsional ketergantungan di antara parameter dalam urutan waktu, untuk menganalisis pola sekuensial, regresi, dan urutan serupa.	Selama musim panas, bencana tanah longsor sering terjadi.

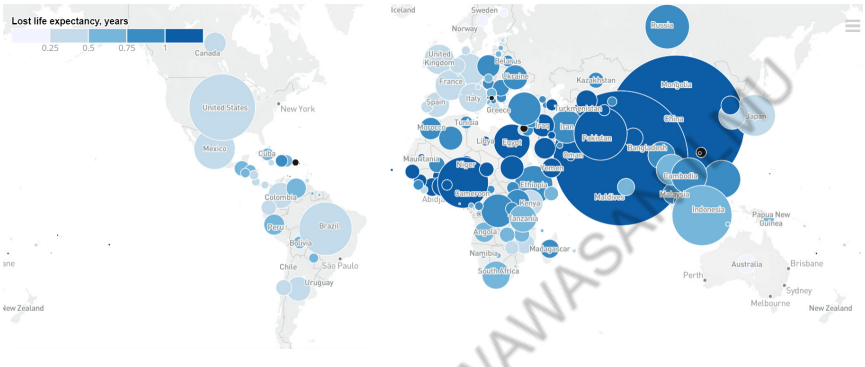
Predictive rule	Tren batin yang meramalkan masa depan nilai dari beberapa variabel ketika pusat temporal atau spasial adalah pindah ke yang lain, atau memprediksi beberapa atribut yang tidak diketahui atau hilang nilai berdasarkan musim lainnya atau informasi berkala.	Prakiraan tren pergerakan longsor berdasarkan pemantauan yang tersedia data. Mengidentifikasi segmen dari populasi yang mungkin merespons mirip dengan acara yang diberikan.
Exception	Pencilan yang diisolasi dari aturan umum atau diturunkan dari pengamatan data lainnya secara substansial, digunakan untuk mengidentifikasi anomaly atribut dan objek.	Deteksi titik pemantauan gerakan luar biasa yang memprediksi tanah longsor atau mendeteksi penipuan transaksi kartu kredit.

Penggunaan spasial data mining juga telah dilakukan secara luas sampai bidang ekonomi. Contohnya Gambar. Memperlihatkan perbandingan pendapatan (dalam) kesetaraan dan pendapatan rata-rata di Swiss. Studi kasus ini dapat diakses di (https://timogrossenbacher.ch/2019/04/bivariate-maps-with-ggplot2-and-sf/) beserta cara pembuatan peta menggunakan software R.



Gambar 10.3 Perbandingan Pendapatan di Swiss

Contoh kasus lain yang telah menggunakan spasial data adalah atlas Sustainable Development Goals (SDG) Gambar. Salah satunya untuk tujuan ke 11 terkait kota dan komunitas berkelanjutan dapat diakses di (https://datatopics.worldbank.org/sdgatlas/goal-11-sustainable-cities-and-communities/). Pada peta dapat dilihat pengurangan umur manusia akibat polusi. Rata-rata pengurangan umur ditunjukkan dengan bubble map sesuai perkirangan level pengurangan umur.



Gambar 10.4 Pengurangan Umur Akibat Polusi

10.4 Studi Kasus

Studi Kasus Data Sensus di United Kingdom

R dapat diunduh dari https://www.r-project.org/ jika belum ada di komputer Anda. Meskipun dimungkinkan untuk melakukan analisis pada R secara langsung, Anda mungkin merasa lebih mudah untuk menjalankannya melalui Rstudio yang menyediakan antarmuka pengguna grafis yang ramah pengguna. Setelah mengunduh R, Rstudio dapat diperoleh secara gratis dari https://www.rstudio.com/

Kasus ini menggunakan data practicaldata.csv . Data berisi Data Sensus area kecil untuk Borough of Camden di UK (Lansley & Cheshire, 2020) . Lebih lengkap data dapat diakses di (https://data.cdrc.ac.uk/dataset/introduction-spatial-data-analysis-and-visualisation-r). Variabel dalam data sensus terdiri dari:

a. Output Area (OA) merupakan data geografis wilayah
b. White British merupakan Suku

c. Low Occupancy merupakan peringkat hunian (kamar tidur) -1 atau kurang
d. Unemployed merupakan aktif secara ekonomi: Pengangguran
e. Qualification merupakan kualifikasi level tertinggi: Kualifikasi level 4 ke atas atau siswa

```
# Set the working directory. Remember to change the example below.
setwd("C:/BIBIB/00_BELAJAR/Belajar R/Spatial Data Mining")
# Load the data. You may need to alter the file directory
Census.Data <-read.csv("practicaldata.csv")
head(Census.Data)
##  OA White_British Low_Occupancy Unemployed Qualification
## 1 E00004120    42.35669   6.2937063  1.893939    73.62637
## 2 E00004121    47.20000   5.9322034  2.688172    69.90291
## 3 E00004122    40.67797   2.9126214  1.212121    67.58242
## 4 E00004123    49.66216   0.9259259  2.803738    60.77586
## 5 E00004124    51.13636   2.0000000  3.816794    65.98639
##  6 E00004125           41.41791         3.9325843
3.846154       74.20635
Instal packages berikut
rgdal - Bindings for the Geospatial Data Abstraction Library
rgeos - Interface to Geometry Engine - Open Source
##Membuat maps di R
# Load packages library ("rgdal")
```

Selanjutnya, kita perlu memuat shapefile area output ke R.

Shapefile terdiri dari beberapa file berbeda yang bila digabungkan dengan paket perangkat lunak tertentu, dapat dipetakan menggunakan sistem proyeksi umum. Kita akan menggunakan batas area output London. Data akan divisualisasikan sebagai file poligon di mana setiap poligon individu mewakili garis besar area keluaran unik dari area yang akan dipelajari. Dalam contoh ini, file data spasial kita dapat ditemukan di folder shapefile dari paket data yang didonload. Beberapa file terdiri dari:

Camden_oa11.dbf

Camden_oa11.prj

Camden_oa11.shp

Camden_oa11.shx

```
# Memanggil data output area shapefiles
Output.Areas<- readOGR(".", "Camden_oa11")
## Warning in OGRSpatialRef(dsn, layer, morphFromESRI = morphFromESRI, dumpSRS =
## dumpSRS, : Discarded datum OSGB_1936 in Proj4 definition: +proj=tmerc +lat_0=49
## +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100000 +ellps=airy +units=m +no_defs
## OGR data source with driver: ESRI Shapefile
## Source: "C:\BIBIB\00_BELAJAR\Belajar R\Spatial Data Mining", layer: "Camden_oa11"
## with 749 features
## It has 1 fields
```

Shapefile yang dimasukkan harus berisi jumlah fitur yang sama dengan jumlah observasi dalam file Data Sensus. Kita sekarang dapat menjelajahi shapefile. Pertama, kita akan memplotnya sebagai peta untuk melihat dimensi spasial dari shapefile. Ini sangat sederhana, cukup masukkan nama objek shapefile ke dalam fungsi plot() standar.

```
# plots the shapefile plot(Output.Areas)
```

#Menggabungkan data

Sekarang kita perlu menggabungkan Data Sensus kita ke shapefile sehingga atribut sensus dapat dipetakan. Data sensus berisi nama unik dari masing-masing area output, ini dapat digunakan sebagai kunci untuk menggabungkan data ke file area output. Kita akan menggunakan fungsi merge() untuk menggabungkan data.

Perhatikan bahwa kali ini kolom untuk nama area output tidak identik, meskipun faktanya berisi data yang sama. Oleh karena

itu, kita harus menggunakan by.x dan by.ya agar fungsi merge menggunakan kolom yang benar untuk menggabungkan data.

```
# joins data to the shapefile
OA.Census <- merge(Output.Areas, Census.Data,
by.x="OA11CD", by.y="OA")
#Mengatur sistem koordinat
```

Penting juga untuk mengatur sistem koordinat, terutama jika Anda ingin memetakan beberapa file yang berbeda. Fungsi *proj4string()* dan *CRS()* memungkinkan kita untuk mengatur sistem koordinat shapefile ke sistem yang telah ditentukan pilihan kita. Sebagian besar data dari Inggris diproyeksikan menggunakan *British National Grid (EPSG:27700)* yang dihasilkan oleh Ordnance Survey, ini termasuk geografi statistik standar yang dihasilkan oleh Office for National Statistics.

Dalam hal ini, shapefile yang awalnya kita unduh dari situs CDRC Data sudah memiliki sistem proyeksi yang benar sehingga kita tidak perlu menjalankan langkah ini untuk objek OA.Census kita. Namun, perlu diperhatikan langkah ini untuk referensi di masa mendatang.

```
# sets the coordinate system to the British National Grid
proj4string(OA.Census) <- CRS("+init=EPSG:27700")
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj =
## prefer_proj): Discarded datum OSGB_1936 in Proj4 definition
## Warning in proj4string(obj): CRS object has comment, which is lost in output
```

#Memetakan data dalam R Sementara fungsi plot cukup terbatas dalam bentuk dasarnya. Beberapa paket memungkinkan kita untuk memetakan data dengan relatif mudah. Mereka juga menyediakan sejumlah fungsi untuk memungkinkan kita menyesuaikan dan mengubah grafik. ggplot2, misalnya, memungkinkan Anda untuk memetakan data spasial. Namun, mungkin yang paling mudah digunakan adalah fungsi-fungsi di dalam pustaka tmap.

Instal packages :

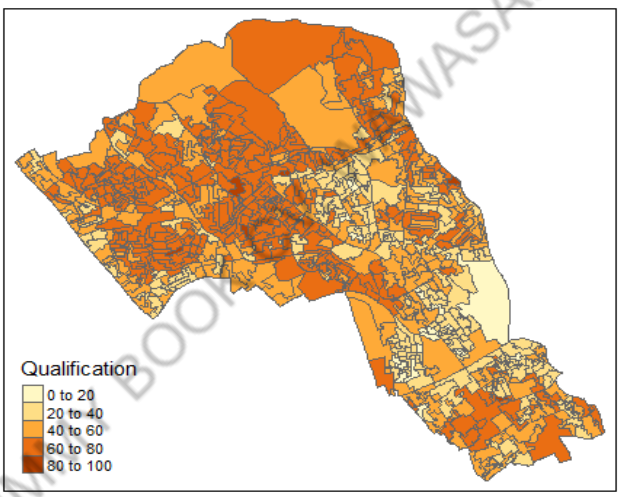
a. ggmap: memperluas paket ploting ggplot2 untuk peta
b. maptools: menyediakan berbagai fungsi pemetaan
c. dplyr dan tidyr: paket manipulasi data yang cepat dan ringkas
d. tmap: paket baru untuk membuat peta yang indah dengan cepat

e. leaflet : salah satu perpustakaan JavaScript open-source paling populer untuk peta interaktif

```
# loads packages library(tmap)
## Warning: package 'tmap' was built under R version 4.1.3
library(leaflet)
## Warning: package 'leaflet' was built under R version 4.1.3
```

#Membuat peta cepat Jika Anda hanya ingin membuat peta dengan legenda dengan cepat, Anda dapat menggunakan fungsi qtm() .

 # this will prodyce a quick map of our qualification variable
qtm(OA.Census, fill = "Qualification")



Membuat peta yang lebih canggih dengan tmap

Membuat peta di tmap melibatkan Anda mengikat beberapa fungsi yang terdiri dari berbagai aspek grafik. Contohnya:

polygon + polygon's symbology + borders + layout

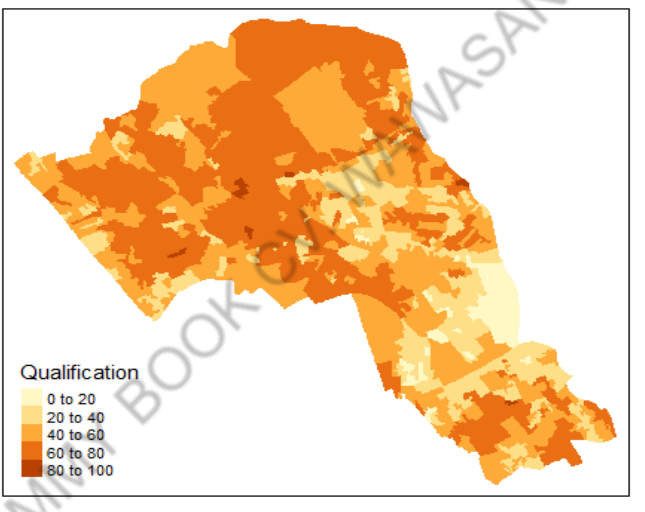
Kami memasukkan shapefile (atau objek data spasial R) diikuti dengan perintah untuk mengatur simbologinya. Objek-objek tersebut kemudian dilapis dalam visualisasi secara berurutan. Objek yang dimasukkan pertama kali muncul di bagian bawah grafik.

#Membuat peta sederhana Di sini kita memuat shapefile dengan fungsi tm_shape() kemudian menambahkan fungsi tm_fill() dimana

kita dapat memutuskan bagaimana poligon diisi. Yang perlu kita lakukan dalam fungsi tm_shape() adalah memanggil shapefile. Kami kemudian menambahkan (+) fungsi baru yang terpisah (tm_fill()) untuk memasukkan parameter yang akan menentukan bagaimana poligon diisi dalam grafik. Secara default, kita hanya perlu memasukkan nama variabel di dalam parameter.

Anda dapat menjelajahi semua fungsi yang tersedia dan bagaimana mereka dapat disesuaikan dengan mengunjungi halaman web untuk tmap, atau dengan memasukkan ? diikuti dengan nama fungsi menjadi r (yaitu ?tm_fill). Di bawah ini kita akan melalui beberapa langkah dasar pemetaan dengan tmap.

# Creates a simple choropleth map of our qualification variable
tm_shape(OA.Census) + tm_fill("Qualification")



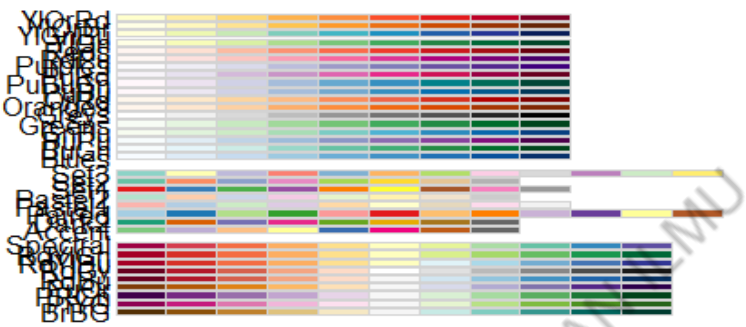
Anda akan melihat bahwa peta sangat mirip dengan peta cepat yang dihasilkan dari penggunaan fungsi qtm(). Namun, keuntungan menggunakan fungsi lanjutan dari tmap adalah mereka menyediakan berbagai opsi penyesuaian. Langkah-langkah berikut akan menunjukkan hal ini.

Mengatur palet warna

tmap memungkinkan Anda untuk menggunakan jalur warna yang ditentukan oleh pengguna atau satu set jalur warna yang telah ditentukan sebelumnya dari fungsi RColorBrewer().

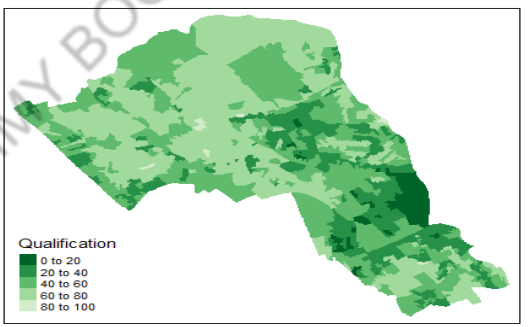
Untuk menjelajahi landai warna yang telah ditentukan di ColourBrewer masukkan kode berikut

```
library(RColorBrewer)
## Warning: package 'RColorBrewer' was built under
R version 4.1.3
display.brewer.all()
```



Ini menyajikan berbagai landai warna yang ditentukan sebelumnya (di atas). Landai terus menerus di bagian atas semuanya sesuai untuk data kami. Jika Anda memasukkan tanda minus sebelum nama jalur di dalam tanda kurung (yaitu -Hijau), Anda akan membalikkan urutan jalur warna.

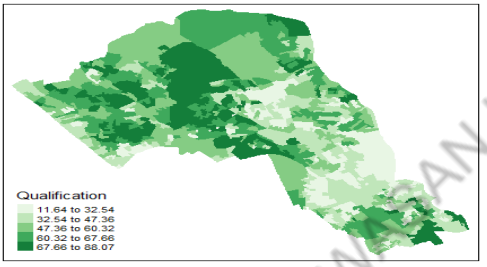
```
# setting a colour palette library(tmap)
tm_shape(OA.Census) + tm_fill("Qualification", pal-
ette = "-Greens")
```



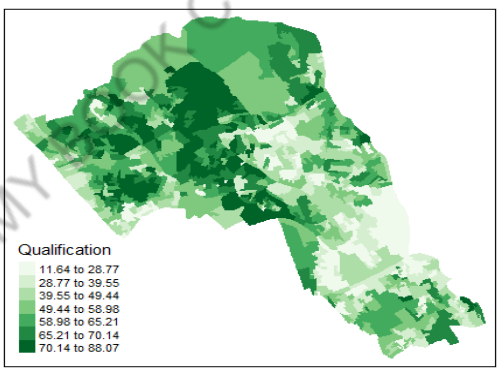
#Mengatur interval warna Kami memiliki berbagai opsi interval yang berbeda dalam parameter gaya. Masing-masing akan sangat memengaruhi cara data Anda divisualisasikan. Untuk melakukan ini, Anda memasukkan "style =" diikuti dengan salah satu opsi di bawah ini.

equal - membagi rentang variabel menjadi n bagian. pretty - memilih sejumlah jeda agar sesuai dengan urutan nilai 'bulat' yang berjarak sama. Jadi kunci mereka untuk interval ini selalu rapi dan mudah diingat. quantile - jumlah kasus yang sama di setiap grup jenks - mencari jeda alami dalam data Cat - jika variabelnya kategorikal (yaitu bukan data kontinu)

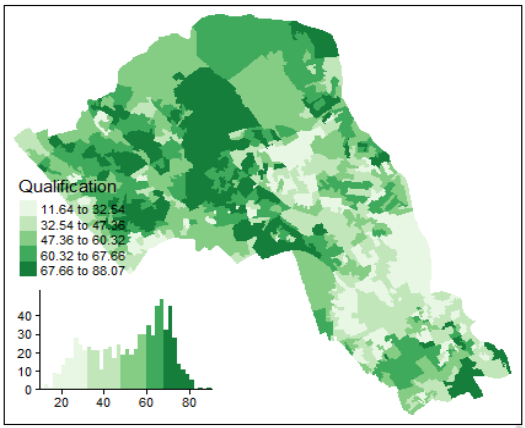
```
# changing the intervals tm_shape(OA.Census) + tm_
fill("Qualification", style = "quantile", palette =
"Greens")
```



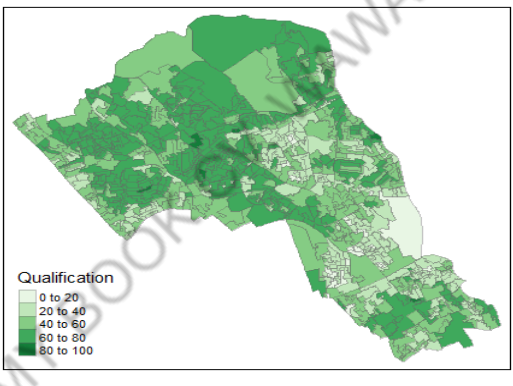

```
# number of levels tm_shape(OA.Census) + tm_
fill("Qualification", style = "quantile", n = 7,
palette = "Greens")
```




```
# includes a histogram in the legend
tm_shape(OA.Census) + tm_fill("Qualification", style
= "quantile", n = 5, palette = "Greens", legend.
hist = TRUE)
```

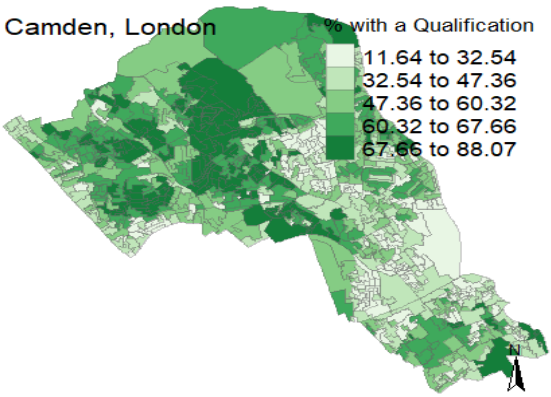



```
# add borders tm_shape(OA.Census) + tm_
fill("Qualification", palette = "Greens") +
tm_borders(alpha=.4)
```



#Mengedit tata letak peta Dimungkinkan untuk mengedit tata letak menggunakan fungsi tm_layout() . Dalam contoh di bawah ini kami juga menambahkan beberapa perintah lagi dalam fungsi tm_fill().

```
# adds in layout, gets rid of frame tm_shape(OA.
Census) + tm_fill("Qualification", palette = "Greens",
style = "quantile", title = "% with a Qualification")
+ tm_borders(alpha=.4) + tm_compass() +
tm_layout(title = "Camden, London", legend.text.size
= 1.1, legend.title.size = 1.4, legend.position =
c("right", "top"), frame = FALSE)
```

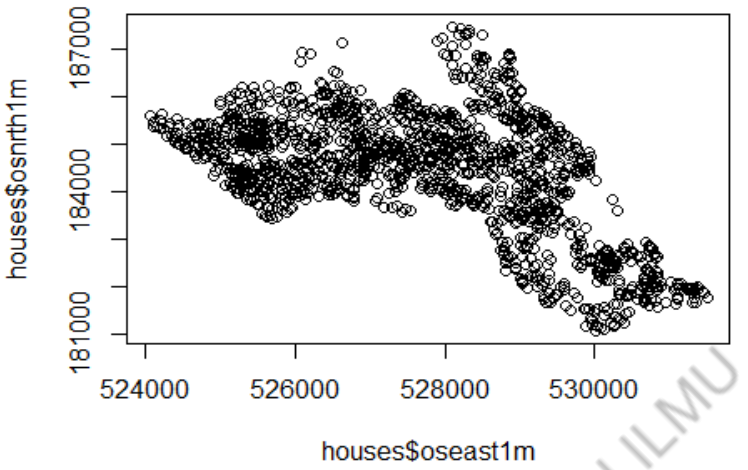



```
#Menyimpan shapefile Terakhir, kita dapat menyimpan
shapefile dengan data Sensus yang sudah dilampirkan
hanya dengan menjalankan kode berikut (ingat untuk
mengubah dsn ke direktori kerja Anda).
#library("rgeos") #writeOGR(OA.Census, dsn = "C:/
BIBIB/00_BELAJAR/Belajar R/Spatial Data Mining",
layer =  "Census_OtA_Shapefile",driver="ESRI
Shapefile3")
```

Memuat data titik ke R

Dalam tutorial ini kita akan menangani data harga rumah berbayar yang awalnya disediakan secara gratis oleh Land Registry. Data diformat sebagai CSV di mana setiap baris adalah penjualan rumah yang unik, termasuk harga yang dibayarkan dalam pound dan kode pos. Sebelum praktik ini, file data digabungkan ke tabel pencarian kode pos Kantor Statistik Nasional (ONS) yang menyediakan koordinat lintang dan bujur untuk setiap kode pos.

```
# load the house prices csv filehouses
<- read.csv("CamdenHouseSales15.csv")
# we only need a few columns for this practical
houses <- houses[,c(1,2,8,9)]
# 2D scatter plot plot(houses$oseast1m,
houses$osnrth1m)
```



Oleh karena itu, kita perlu menetapkan atribut spasial ke CSV sehingga dapat dipetakan dengan benar di R. Untuk melakukan ini kita perlu memuat paket sp, paket ini menyediakan kelas dan metode untuk menangani data spasial. Ingatlah untuk menginstal paket terlebih dahulu jika Anda belum melakukannya sebelumnya.

Selanjutnya, kita akan mengubah CSV menjadi SpatialPointsDataFrame. Untuk melakukan ini, kita perlu mengatur data apa yang akan dimasukkan, kolom apa yang berisi koordinat x dan y, dan sistem proyeksi apa yang kita gunakan.

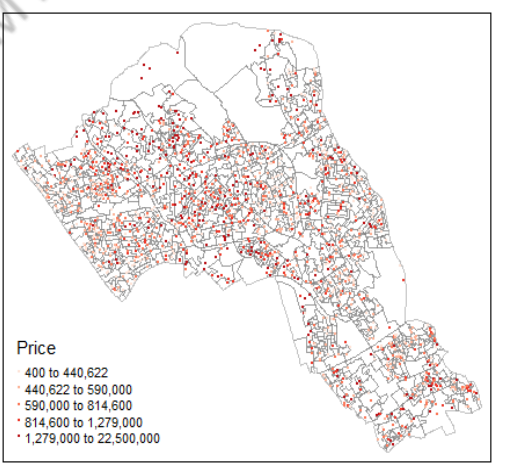
```
library("sp")
# create a House.Points SpatialPointsDataFrame
House.Points <-SpatialPointsDataFrame(houses[,3:4], houses, proj4string = CRS("+init=EPSG:27700"))
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj =
## prefer_proj): Discarded datum OSGB_1936 in Proj4 definition
library("tmap")
# This plots a blank base map, we have set the transparency of the borders to 0.4
tm_shape(OA.Census) + tm_borders(alpha=.4)
```

Kita sekarang dapat menambahkan titik sebagai lapisan tm_ shape tambahan di peta kita.

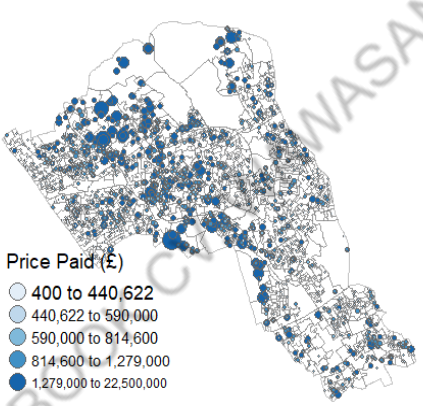
Untuk melakukan ini, kami menyalin kode yang sama untuk membuat peta dasar kami, diikuti dengan simbol plus, lalu masukkan detail untuk data poin. Argumen tambahan untuk data poin dapat diringkas sebagai:

tm_shape(polygon file) + tm_borders(transparency = 40%) + tm_shape(our spatial points data frame) + tm_dots (what variable is coloured, the colour palette and interval style)

```
# creates a coloured dot map
tm_shape(OA.Census) + tm_borders(alpha=.4) +
tm_shape(House.Points) + tm_dots(col = "Price",
palette = "Reds", style = "quantile")
```



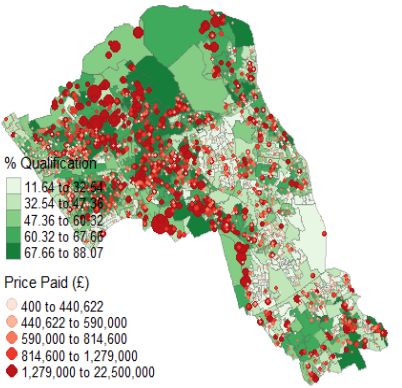
```
#Simbol yang proporsional
# creates a proportional symbol map
tm_shape(OA.Census) + tm_borders(alpha=.4) +
tm_shape(House.Points) + tm_bubbles(-
size = "Price", col = "Price", palette =
"Blues", style = "quantile", legend.size.
show = FALSE, title.col = "Price Paid (£)") +
tm_layout(legend.text.size = 1.1, legend.title.
size = 1.4, frame = FALSE)
## Some legend labels were too wide. These labels
have been resized to 0.92, 0.92, 0.84, 0.74. Increase
legend.width (argument of tm_layout) to make the
legend wider and therefore the labels larger.
```



Kita juga dapat menambahkan lebih banyak argumen dalam fungsi tm_dots() untuk titik seperti yang kita lakukan dengan tm_fill() untuk data poligon. Beberapa argumen unik untuk tm_dots(). Misalnya, argumen skala yang mengubah ukuran titik.

```
# creates a proportional symbol map
tm_shape(OA.Census) + tm_fill("Qualification", palette
= "Greens", style = "quantile", title = "% Qualification") +
tm_borders(alpha=.4) +
tm_shape(House.Points) + tm_bubbles(size =
"Price", col = "Price", palette = "Reds", style
= "quantile", legend.size.show = FALSE, title.
col = "Price Paid (£)", border.col = "black",
border.lwd = 0.1, border.alpha = 0.1) +
tm_layout(legend.text.size = 0.8, legend.title.size
```

```
= 1.1, frame = FALSE)
```



Simpan hasil peta yang diperoleh dengan perintah di bawah ini.

```
# write the shapefile to your computer (remem-
ber to chang the dsn to your workspace)
writeOGR(House.Points, dsn = "C:/BIBIB/00_BELAJAR/
Belajar R/Spatial Data Mining", layer = "Camden_
house_sales", driver="ESRI Shapefile2")
```


10.4 Latihan Soal

1. Jelaskan Jelaskan contoh jurnal penerapan Spatial Data Mining dengan judul Spasial Data Mining Menggunakan Model Sar – Kriging pada link berikut ini https://jurnal.ugm.ac.id/ijccs/article/view/5213/4267

BAB XI
WEB MINING

11.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk

- Mengenalkan tentang web mining
- Mengenalkan web mining framework
- Metode dalam web mining

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai pemahaman mengenai web mining dan aplikasinya.

11.2 Web Mining

Web mining adalah Penambangan web adalah penggunaan teknik penambangan data untuk secara otomatis menemukan dan mengekstrak informasi dari dokumen/layanan Web (Etzioni, 1996, CACM 39).

Web Mining bertujuan untuk penemuan yang bermanfaat informasi atau pengetahuan dari Web struktur hyperlink, konten halaman dan data penggunaan (Bing LIU 2007, Web Data Mining, Springer)

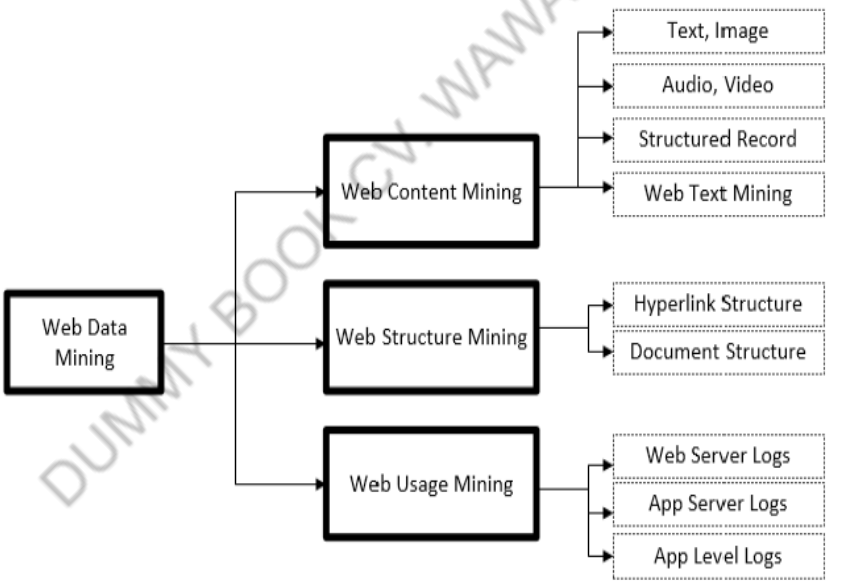
Penambangan data web menjadi platform yang mudah dan penting untuk pengambilan informasi yang berguna. Pengguna lebih memilih World Wide Web untuk mengunggah dan mengunduh data. Seiring meningkatnya pertumbuhan data melalui internet, semakin sulit dan memakan waktu untuk menemukan pengetahuan dan pola yang informatif. Menggali informasi yang berpengetahuan dan pengguna dari data yang tidak terstruktur dan tidak konsisten melalui web bukanlah tugas yang mudah untuk

dilakukan. Teknik penambangan yang berbeda digunakan untuk mengambil informasi yang relevan dari web (hyperlink, konten, log penggunaan web). Penambangan data web adalah subdisiplin penambangan data yang terutama berhubungan dengan web. Penambangan data web dibagi menjadi tiga jenis berbeda: struktur web, konten web, dan penambangan penggunaan web. Semua jenis ini menggunakan teknik, alat, pendekatan, algoritme yang berbeda untuk menemukan informasi dari sejumlah besar data melalui web.

KATEGORI PERTAMBANGAN WEB

Penambangan Web dikategorikan menjadi tiga jenis yaitu:

A. Penambangan Konten Web
B. Penambangan Struktur Web
C. Penambangan Penggunaan Web



Gambar 11.1 Taxonomy Web Mining (Mughal,2018)

Web Mining terdiri dari masif, dinamis, beragam dan sebagian besar data tidak terstruktur yang menyediakan sejumlah besar data. Pertumbuhan web yang eksplosif menyebabkan beberapa masalah seperti menemukan data yang relevan melalui internet, mengamati perilaku pengguna. Untuk memecahkan masalah semacam itu,

upaya dilakukan untuk menyediakan data yang relevan dalam bentuk struktur (tabel) yang mudah untuk memahami dan berguna bagi organisasi untuk memprediksi kebutuhan pelanggan. (Kshitija Pol K,et. al. 2008)

A. Penambangan Konten Web

Penambangan Konten adalah proses Penambangan Web di mana data informatif yang diperlukan diambil dari situs web (www).

1) Teknik Penambangan Konten Web: Penambangan konten web menggunakan teknik yang berbeda. Berikut ini adalah empat teknik yang dijelaskan digunakan oleh penambangan konten web. Sebagian besar data konten web dalam bentuk teks tidak terstruktur. Untuk ekstraksi data tidak terstruktur, penambangan konten web membutuhkan pendekatan text mining dan data mining (Malarvizhi, Saraswathi,2013). Teks dokumen terkait dengan penambangan teks, pembelajaran mesin, dan bahasa alami. Tujuan utama dari penambangan teks adalah untuk mengekstrak informasi sebelumnya dari sumber konten (Johnson, Gupta,2012). Penambangan teks adalah bagian dari penambangan konten web dan karenanya teknik yang berbeda adalah digunakan untuk penambangan data teks dari konten web melalui internet/website untuk memberikan data yang tidak diketahui, beberapa di antaranya adalah disebutkan di bawah:
   - Ekstraksi Informasi
   - Ringkasan
   - Visualisasi Informasi
   - Pelacakan Topik
   - Kategorisasi
   - Pengelompokan

   Terstruktur adalah teknik yang menambang data terstruktur di web. Data mining struktur adalah teknik penting karena itu mewakili halaman host di web. Dibandingkan dengan tidak terstruktur, dalam penambangan data terstruktur selalu mudah untuk mengekstrak data (Herrouz et. al.2012). Berikut ini adalah beberapa teknik yang digunakan untuk penambangan data terstruktur:
   - Web Content Mining

- Web Structure Mining
- Web Usage Mining

Semi-terstruktur adalah bentuk data terstruktur tetapi tidak lengkap, teks dalam data semi-terstruktur adalah tata bahasa. Strukturnya Hierarkis, tidak ditentukan sebelumnya. Representasi data semi terstruktur dalam bentuk tag (seperti HTML, XML). HTML adalah kasus struktur intra-dokumen [4]. Teknik yang digunakan untuk mengekstrak data semi terstruktur adalah:

- OEM - Model Pertukaran Objek
- Bahasa Ekstraksi Data Web
- Ekstraksi Top Down (Malarvizhi, Saraswathi,2013)

Secara tradisional komputasi data dianggap sebagai teks dan angka, tetapi sekarang ada berbagai jenis data komputasi yang berbeda dari data multimedia seperti video, gambar, audio, dll. Proses penambangan ini digunakan untuk mengekstrak kumpulan data multimedia yang menarik dan juga mengubah jenis kumpulan data menjadi digital. media (Vijayarani, Sakila,2015). Teknik yang digunakan untuk data mining multimedia adalah:

- Penambang Multimedia
- Deteksi Batas Tembakan
- SKICAT
- Pencocokan Histogram Warna (Kumar, Singh,2017)

2) Algoritma Penambangan Konten Web: Beberapa teknik digunakan oleh penambangan web untuk mengekstrak informasi dari sejumlah besar basis data. Ada berbagai jenis algoritma yang digunakan untuk mengambil informasi pengetahuan, di bawah ini dijelaskan beberapa algoritma klasifikasi:

Pohon keputusan merupakan pendekatan berbasis klasifikasi dan terstruktur yang terdiri dari simpul akar, cabang dan simpul daun. Ini adalah proses hierarkis di mana simpul akar dipecah menjadi sub cabang dan simpul daun berisi label kelas.

Pohon keputusan adalah teknik yang kuat.

Naïve Bayes adalah algoritma yang mudah, sederhana, kuat untuk klasifikasi dan juga dikenal sebagai classifier Native Bayes. Ini didasarkan pada Teorema Bayes. Dari nilai set data yang telah ditentukan, probabilitas dihitung untuk setiap kelas dengan menghitung kombinasi nilai. Kelas yang paling mungkin adalah kelas dengan probabilitas tertinggi (Patil, Sherekar,2013).

Teorema Bayes: (Bilal et. Al,2013) Support Vector Machine adalah algoritma klasifikasi dan pembelajaran mesin yang terkenal dan sederhana. SVM merupakan metode yang dapat digunakan untuk kumpulan data linier dan nonlinier (Kumar, Singh,2017). Hyper plane pemisah yang optimal (decision boundary) hanyalah sebuah garis yang digunakan untuk menggambar untuk memisahkan dua kelas tergantung pada fitur klasifikasi yang berbeda.

Jaringan saraf tiruan adalah pendekatan penambangan konten web lain yang menggunakan algoritma propagasi balik. Algoritma ini terdiri dari beberapa lapisan yaitu lapisan input, beberapa lapisan tersembunyi dan kemudian lapisan output, masing-masing mengumpankan lapisan berikutnya sampai lapisan terakhir (output). Neuron adalah unit dasar dari jaringan saraf. Input diumpankan secara bersamaan ke unit. Dari lapisan input, input secara bersamaan diumpankan ke lapisan tersembunyi. Biasanya ada satu lapisan tersembunyi tetapi jumlah lapisan tersembunyi berubah-ubah [10]. Lapisan tersembunyi terakhir memberi makan input dan membentuk lapisan output.

Selengkapnya tentang teks sumber ini diperlukan teks sumber untuk mendapatkan informasi terjemahan tambahan

Kirim masukan

Panel samping

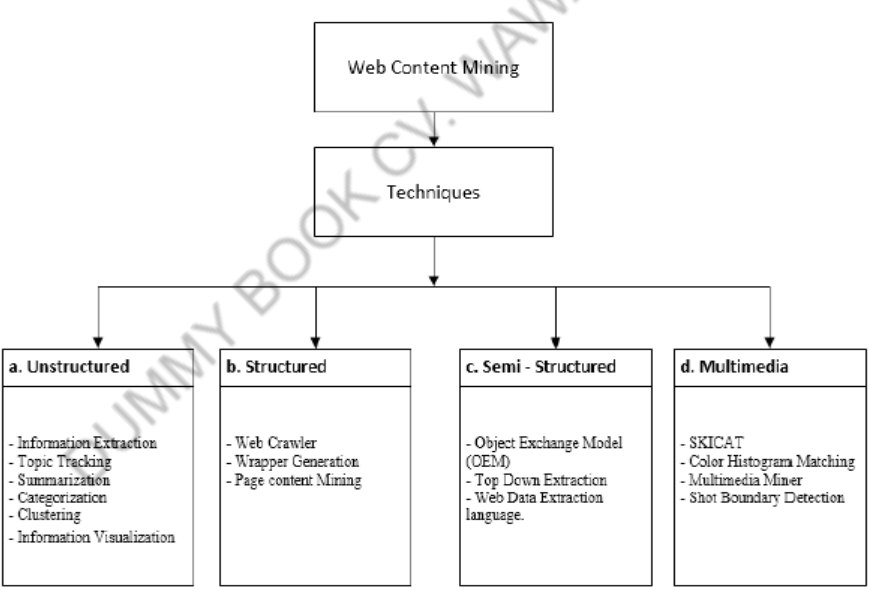
Tabel 11.1 Ringkasan Kategori Web Mining (Mughal,2018)

Web Mining Categories	Techniques	Tools	Algorithms
Web Content Mining	- Unstructured Data Mining - Structured Data Mining - Semi – Structure Data Mining - Multimedia Data Mining	- Screen Scaper - Mozenda - Automation Anywhere7 - Web Content Extractor - Web Info Extractor - Rapid Miner	- Decision Tree - Naive Bayes - Support Vector Machine - Neural Network
Web Structure Mining	- Link-based Classification - Link-based Cluster Analysis - Link Type - Link Strength - Link Cardinality	- Google PR Checker - Link Viewer	- Page Rank Algorithm - HITS algorithms (Hyperlink Induced Topic Search) - Weighted Page Rank Algorithm - Distance Rank Algorithm - Weighted Page Content Rank Algorithm - Webpage Ranking Using Link Attributes - Eigen Rumor Algorithm - Time Rank Algorithm -Tag Rank Algorithm - Query Dependent Ranking Algorithm
Web Usage Mining	- Data Preprocessing • Data Cleaning • User & Session Identification - Pattern Discovery • Statistical Analysis • Association Rules • Clustering • Classification • Sequential Patterns - Pattern Analysis • Knowledge Query Mechanism • OLAP (Online Analytical processing) • Intelligent Agents	- Data Preprocessing Tools • Data Preparator • Sumatra TT • Lisp Miner • SpeedTracer - Pattern Discovery Tools • SEWEBAR-CMS • i-Miner • Argunaut • MiDas(Mining In-ternet Data for As-sociative Sequenc-es) - Pattern Analysis Tools • Webalizer • Naviz • WebViz • WebMiner • Stratdyn	- Association Rules • Apriori Algorithm • Maxi-mal Forward References • Markov Chains • FP Growth • Prefix Span - Clustering • Self-Organized Maps • Graph Partitioning • Ant Based Technique • K-means with Genetic algorithms • Fuzzy c-mean Algorithm - Classification • Decision Trees • Naïve Bayesian Classifiers • K-nearest Neighbor Classifiers • Support Vector Machine - Sequential Patterns • MIDAS (Mining Internet Data for Association Sequences) algorithm

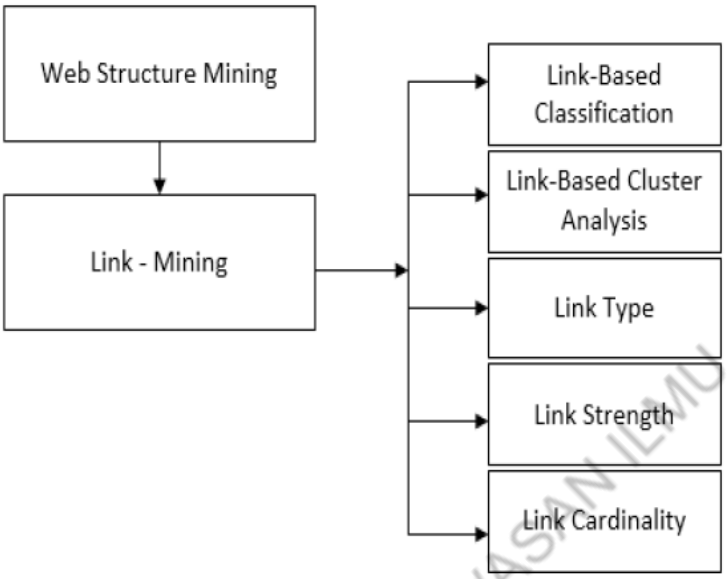
Alat Web Usage Mining:

Speed Tracer adalah alat analisis dan penggunaan untuk penambangan penggunaan. Alat ini membantu menemukan perilaku berselancar pengguna dan menganalisis dengan entri yang disimpan di log server dengan menggunakan teknik penambangan data yang berbeda. Cookie tidak diperlukan untuk mengidentifikasi sesi pengguna, pelacak kecepatan menggunakan berbagai jenis informasi seperti: alamat IP, URL halaman, agen, dll. Kumpulan pola penelusuran membantu memahami perilaku pengguna dengan cara yang lebih baik (Pranit, Chawan,2013). Tiga jenis pemahaman yang dihasilkan oleh pelacak kecepatan: Berbasis pengguna yang mengacu pada durasi waktu akses pengguna. Path based berhubungan dengan proses path yang sering dikunjungi di web. Berbasis grup

menghasilkan informasi grup halaman web yang dikunjungi berulang kali. Suggest 3.0 adalah sistem yang menyediakan informasi yang familiar bagi pengguna tentang halaman web yang mungkin mereka minati. Kebutuhan pelanggan/pengguna berhasil dicapai dengan serangkaian perubahan konstan pada tautan halaman. Sarankan 3.0 menggunakan algoritme partisi grafik untuk mempertahankan informasi historis. Tujuan utamanya adalah untuk mencatat perubahan inkremental (Pranit,Chawan,2013). WebViz adalah alat yang digunakan untuk analisis statistik log akses web. Ide utama dari pengembangan alat ini adalah untuk memberikan tampilan grafis kepada perancang basis data WWW dari db lokal dan pola akses mereka. Hubungan antara log akses dan database (lokal) ditampilkan dengan menggunakan paradigma Web-Path (Pitkow, Bharat,1994). Ini menyajikan dokumen database lokal dan asosiasi dokumen dalam struktur grafik. Informasi tentang dokumen yang diakses dikumpulkan dari log akses. Jumlah jalur yang dikunjungi oleh pengguna juga dikumpulkan untuk ditampilkan.



Gambar 11.2 Teknik Web Mining (Johnson F and Kumar Santosh Gupta K.S, 2012)



Gambar 11.3 Web Structure Mining

1. Klasifikasi Berbasis Tautan: Ini adalah versi klasifikasi pemutakhiran dari penambangan data klasik dan tugasnya adalah menautkan domain. Fokus utama adalah untuk memprediksi kategori halaman web – berdasarkan teks, tag HTML, link antara halaman web dan atribut lainnya (Xiaoguang, Brian, 2009).
2. Analisis Cluster Berbasis Tautan: Fokus utama adalah pada segmentasi data. Dalam analisis klaster data dikategorikan atau dikelompokkan bersama (John,2001). Objek serupa dikelompokkan dalam satu kelompok dan objek data yang berbeda dikelompokkan secara terpisah. Untuk menggali pola tersembunyi dari dataset dapat digunakan analisis cluster berbasis link (Xiaoguang, Brian, 2009).
3. Jenis Tautan: Ini membantu untuk menebak jenis tautan antara entitas (dua atau lebih) (John,2001).
4. Kekuatan Tautan: Kekuatan tautan menunjukkan bahwa tautan mungkin terkait dengan bobot (John,2001).

Perbandingan Penggunaan Teknik Penambangan Data

Usage Mining Techniques	Methods	Data Gathering	Data Store	Advantages	Important Algorithms
Data Preprocessing	- Web status codes	- Data logs - Website - Users login information - Web access logs - Cache - Cookies etc.	- Web logs	- Convert raw data to understandable Common LF and Extended CLF for recording	- Apriori algorithm - FP Growth
Pattern Discovery	- Frequency, median, mode used to show length, recently accessed, view time of pages	- Filtered data from preprocessing section	- Session logs	- Extract useful information from discovered patterns correlations	- K-means with Genetic algorithms - Fuzzy c-mean Algorithm
Pattern Analysis	- Roll-up - Drill Down/Up	- Pattern discovery	- Data cube (multi-dimensional database)	- Irrelevant rules and patterns are separated	- SQL Language - OLAP

Gambar 11.4 Perbandingan penggunakan Teknik Mining (Mughal,2018)

Penambangan data web dianggap sebagai sub pendekatan penambangan data yang berfokus pada pengumpulan informasi dari web. Web adalah domain besar yang berisi data dalam berbagai bentuk yaitu: gambar, tabel, teks, video, dll. Karena ukuran web adalah terus meningkat; itu menjadi tugas yang sangat menantang untuk mengekstrak informasi. Penambangan konten web berguna dalam istilah mengeksplorasi data dari teks, tabel, gambar, dll. Web penambangan struktur mengklasifikasikan hubungan antara halaman web yang ditautkan. Penambangan penggunaan web juga merupakan jenis penting yang menyimpan pengguna mengakses data dan mendapatkan informasi tentang pengguna tertentu dari log. Semua teknik mungkin memiliki beberapa kelebihan dan kekurangan tapi kekurangannya bisa diperbaiki dengan studi lebih lanjut.

11.3 Contoh Soal

1. Jelaskan beberapa teknik yang digunakan untuk penambangan data terstruktur!
2. Jelaskan web structure mining!

BAB XII
TEXT MINING

12.1 Tujuan dan Capaian Pembelajaran

Tujuan:

Bab ini bertujuan untuk

- Mengenalkan tentang text mining
- Mengenalkan text mining framework
- Memberikan Studi Kasus Text Mining

Capaian Pembelajaran:

Capaian pembelajaran mata kuliah pada bab ini adalah mahasiswa mampu menguasai pemahaman mengenai text mining dan aplikasinya.

12.2 Pendahuluan

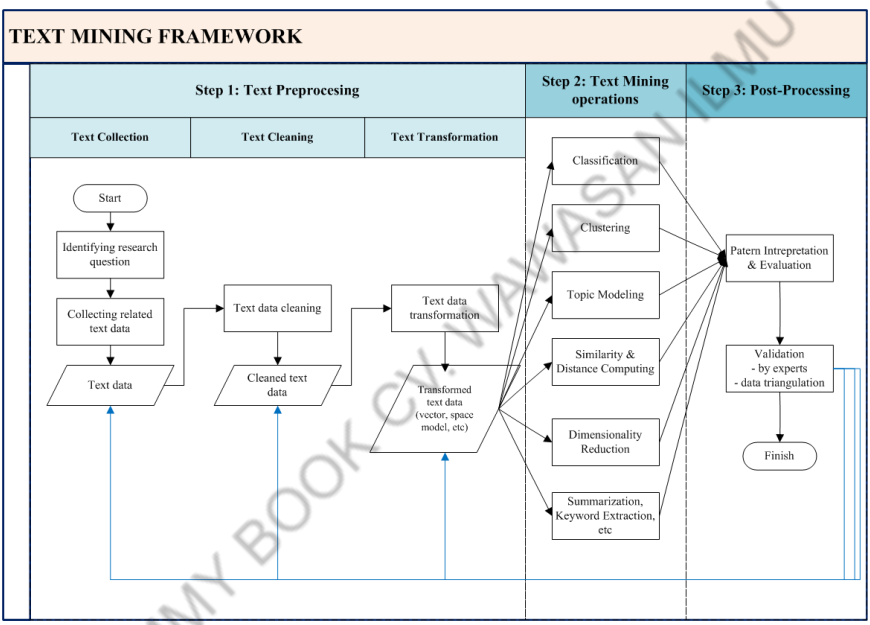
Data yang dimiliki oleh suatu industri tidak hanya data kuantitatif namun kebanyakan berupa data tekstual atau kualitatif. Hal ini karena perusahaan biasa mencatat peristiwa atau laporan misalnya dalam pengecekan kualitas produk dengan memberikan catatan. Data berupa teks merupakan data yang tidak terstruktur dan perlu perlakuan khusus agar dapat ditarik kesimpulan berbeda dengan data terstruktur yang dapat menggunakan statistik deskriptif maupun data mining. Data mining dan text mining berbeda pada jenis data yang mereka tangani. Data mining dapat menangani data terstruktur yang berasal dari sistem, seperti database, spreadsheet, *enterprise resource planning*, *customer relationship management*, dan aplikasi akuntansi, sedangkan text mining berurusan dengan data tidak terstruktur yang ditemukan dalam dokumen, email, media sosial, atau website. Jadi, perbedaan antara data mining dan text mining adalah bahwa dalam text mining, pola diekstraksi dari teks bahasa alami bukan dari database fakta terstruktur. Semua informasi tertulis atau lisan yang dapat seringkali ditemukan di

sebuah industri dapat dipresentasikan secara tekstual dan dianalisis menggunakan text mining. Bab ini akan membahas tentang text mining dan aplikasinya dalam studi kasus teknik industri.

12.3 Text Mining

Sejumlah besar informasi tersedia dalam ekonomi digital saat ini berupa data tekstual yang lebih mudah jika diklasifikasikan atau dikategorikan ke dalam beberapa kelas yang telah ditentukan. Seperti pada lingkungan bisnis atau industri apa pun, informasi perusahaan tersedia dalam berbagai format, sekitar 80% di antaranya adalah dokumen teks. Informasi ini dalam bentuk data deskriptif dalam format dan mencakup laporan layanan informasi perbaikan, dokumen kualitas manufaktur, dan catatan meja bantuan pelanggan. Ini juga sering dalam format teks ringkas dan berisi banyak istilah dan singkatan khusus industri. Baik upaya teknis maupun manual diperlukan untuk memproses sumber-sumber ini, menyelidiki pola-pola, dan menemukan pengetahuan berguna yang tersembunyi di dalam sumber-sumber ini. Mengubah sumber yang berguna ini menjadi format yang dapat digunakan akan membantu meningkatkan kualitas produk atau layanan di masa mendatang dan memberikan solusi untuk masalah manajemen proyek. Oleh karena itu, dengan menemukan pola pengetahuan yang berguna, Anda dapat mendukung pembuat keputusan atau pekerja pengetahuan dan meningkatkan keputusan bisnis Anda. Pengetahuan yang teridentifikasi juga dapat ditransfer dari satu proyek ke proyek lainnya. Pada akhirnya, ini membantu meningkatkan kualitas produk atau layanan Anda dan mengurangi upaya yang terkait dengan manajemen proyek. Oleh karena itu, tujuan dari penelitian ini adalah untuk mencoba meningkatkan konteks bisnis di mana pelajaran yang bermanfaat dari pengalaman sebelumnya dimuat dalam laporan dan dokumen lainnya. Misalnya, mampu mengidentifikasi dan mengkategorikan kebutuhan pelanggan akan meningkatkan pengambilan keputusan di masa depan dan meningkatkan kepuasan pelanggan. Jika data mining dalam proses keseluruhan penemuan pengetahuan dalam database (*Knowledge Discovery in Database*/KDD) adalah identifikasi pola data yang valid, baru, berpotensi berguna, dan pada akhirnya dapat dipahami. Maka pada text mining istilah penemuan pengetahuan dari database tekstual (*Knowledge Discovery in Textual Database*/KDT) sedikit

berbeda dengan istilah bentuk umum KDD dan dapat didefinisikan sebagai menemukan informasi yang berguna dan pengetahuan dari database tekstual melalui penerapan teknik data mining. Data mining maupun text mining berbagi metode umum untuk mengumpulkan informasi dari data mentah dan mengolahnya melalui penerapan teknik data mining. Sebelum teknik data mining dapat diterapkan maka harus dilakukan tida proses dalam text mining yaitu yaitu pengumpulan data, pra-pemrosesan dan aplikasi teknik penambangan teks diperlukan. Secara lengkap alur proses text mining dapat dilihat pada Gambar.



Gambar 12.1 Text Mining Framework diadaptasi dari ((Kobayashi et al., 2018))

12.4 Langkah-langkah Text Mining

Kerangka kerja text mining dibagi menjadi tiga langkah utama yaitu dimulai dari preproses data teks, operasi text mining sampai post proses dimana hasil analisis akan diinterpretasi. Setiap langkah utama masih terbagi menjadi tahapan-tahapan kecil. Misalnya dalam text pre-proses masih akan dibagi menjadi pengumpulan teks, pembersihan teks dan transformasi teks. Secara lengkap tahapan dapat dijelaskan di bawah:

Langkah 1. Text preprocessing

a. Pengumpulan data teks (*text data collection*)

Sebelum memulai proses text mining, seseorang harus memiliki data teks dan langkah pertama dalam pengumpulan data adalah memutuskan sumber data yang paling sesuai. Sumber potensial termasuk web, dokumen perusahaan (misalnya, memo, laporan, dan penawaran perekrutan), teks pribadi (misalnya, buku harian, email, pesan SMS, dan tweet; dan tanggapan survei terbuka. Text mining mengharuskan teks harus dalam bentuk digital atau dapat ditranskripsikan ke formulir ini. Jika tidak, teks nondigital (misalnya, dokumen tulisan tangan atau cetakan) dapat didigitalkan menggunakan karakter optik teknik pengenalan. Data teks web dikumpulkan dari situs web baik melalui antarmuka pemrograman aplikasi web (*applications programming interfaces*/API) atau *web scraping* yaitu, ekstraksi otomatis web konten halaman.

Penting untuk menyadari masalah hukum dan etika yang terkait dengan akses data, khususnya web scraping perlu untuk diperhatikan. Konten situs web sering kali dilindungi oleh undang-undang hak cipta dan tuntutan hukum dapat terjadi jika perjanjian di bawah penggunaan wajar dilanggar. Selain itu juga, masalah privasi mungkin melarang penggunaan jenis data teks pribadi tertentu tanpa izin, seperti formulir web, survei, email, dan penilaian kinerja. Masalah potensial lainnya yang harus diperhatikan adalah penyimpanan data. Dalam proyek kecil, data teks yang dikumpulkan dapat disimpan sementara dalam sistem file lokal (misalnya, di komputer). Namun, dalam analitik teks skala besar, terutama ketika data berasal dari sumber yang berbeda, penggabungan, penyimpanan, dan pengelolaan data mungkin memerlukan integrasi sistem database atau gudang data.

b. Pembersihan data teks (*text data cleaning*)

Pembersihan data meningkatkan kualitas data, yang pada gilirannya meningkatkan validitas pola dan hubungan yang diekstraksi. Pembersihan dilakukan dengan mempertahankan hanya elemen teks yang relevan. Prosedur pembersihan standar untuk teks termasuk penghapusan karakter yang tidak penting (misalnya spasi ekstra, tag pemformatan, dll.), "segmentasi teks", "konversi huruf kecil", "stop penghilangan kata", dan "pembentukan kata". Untuk tanggapan survei terbuka (dan lainnya secara informal teks yang dihasilkan seperti teks SMS atau email

pribadi), menurut pengalaman kami, mungkin berguna untuk menjalankan pemeriksaan ejaan untuk mengoreksi kata-kata yang salah eja. Untuk dokumen web, tag HTML atau XML harus dihapus karena ini tidak menambahkan konten yang berarti. Jadi hasil akhirnya adalah data teks dilucuti dari semua kata dan karakter konten rendah.

Segmentasi teks adalah proses membagi teks menjadi kalimat dan kata. Stopwords seperti konjungsi dan preposisi (misalnya, dan, the, dari, untuk) adalah kata-kata yang memiliki rendah konten informasi dan tidak berkontribusi banyak pada makna dalam teks. "Stemming" menjadi homogen representasi kata-kata yang mirip secara semantik (misalnya, mewakili kata-kata "memastikan," "memastikan," dan "dipastikan" dengan "memastikan"). Karena teknik ini menghapus kata, mereka juga berfungsi untuk mengurangi ukuran kosakata.

c. Transformasi teks (*Text transformation*)

Transformasi teks adalah strategi kuantifikasi di mana teks diubah menjadi struktur matematika. Sebagian besar teknik analisis memerlukan teks untuk diubah menjadi matriks struktur, di mana kolom adalah variabel (juga disebut sebagai fitur) dan baris adalah dokumen. Salah satu cara untuk membangun matriks ini adalah dengan menggunakan kata- kata atau istilah dalam kosakata sebagai variabel. Matriks yang dihasilkan disebut "matriks dokumen demi suku" di mana nilai-nilai variabel adalah "bobot" dari kata-kata dalam dokumen itu. Dalam banyak aplikasi, ini adalah pilihan langsung karena kata-kata adalah unit linguistik dasar yang mengungkapkan makna. Frekuensi mentah dari kata adalah jumlah kata itu dalam dokumen. Jadi dalam transformasi ini, setiap dokumen adalah diubah menjadi "vektor", yang ukurannya sama dengan ukuran kosakata, dengan masing-masing elemen yang mewakili bobot istilah tertentu dalam dokumen itu.

Frekuensi kata itu sendiri mungkin tidak berguna jika tugasnya adalah membuat pengelompokan atau kategori dokumen. Pertimbangkan kata studi dalam kumpulan abstrak artikel ilmiah. Jika tujuannya adalah untuk mengkategorikan artikel ke dalam topik atau tema penelitian maka kata ini tidak informatif karena dalam konteks khusus ini hampir semua dokumen mengandung kata ini. Cara untuk mencegah penyertaan istilah yang memiliki sedikit kekuatan diskriminatif adalah untuk menetapkan bobot untuk

setiap kata sehubungan dengan kekhususannya untuk beberapa dokumen dalam korpus. Prosedur pembobotan yang paling umum digunakan untuk ini adalah kebalikannya frekuensi dokumen (IDF). Sebuah istilah tidak penting dalam diskriminasi proses jika IDF-nya adalah 0, menyiratkan bahwa kata tersebut ada di setiap dokumen. Bahkan, IDF juga bisa dasar untuk memilih kata henti untuk tugas kategorisasi yang ada. Kata-kata yang memiliki IDF rendah memiliki sedikit kekuatan diskriminatif dan dapat dibuang. Jika dikalikan, kata raw frequency (tf) dan IDF menghasilkan TF-IDF yang populer, yang secara bersamaan memperhitungkan pentingnya sebuah kata dan spesifisitasnya.

Mewakili teks sebagai matriks dokumen-demi-istilah mengandaikan bahwa informasi urutan kata tidak krusial dalam analisis. Meskipun tidak canggih, perlu dicatat bahwa transformasi yang mengabaikan informasi urutan kata berkinerja lebih baik di banyak aplikasi daripada transformasi yang menjelaskannya. Tantangan komputasi utama untuk representasi matriks dokumen-demi-istilah adalah bagaimana menangani dimensi yang dihasilkan, yang berbanding lurus dengan ukuran kosakata. Seseorang dapat menggunakan dimensi data yang berbeda metode reduksi untuk mengurangi jumlah variabel (misalnya, pemilihan variabel dan proyeksi variabel teknik) atau menggunakan teknik khusus yang cocok untuk data dengan dimensi tinggi. Teknik-teknik ini akan disorot di bagian Operasi Penambangan Teks. Setelah teks diubah, teknik seperti analisis regresi dan analisis klaster dapat digunakan terapan. Menggabungkan variabel membantu menjawab pertanyaan substantif tentang teks (lihat juga Teks bagian Operasi Pertambangan). Misalnya, jika resume pelamar kerja digunakan sebagai sumber data, maka kehadiran kata-kata pengalaman dan tahun bersama dengan angka dapat digunakan untuk menyimpulkan suatu pengalaman kerja pelamar.

12.5 Operasi Text Mining

Meskipun transformasi teks mendahului penerapan metode analitis, kedua langkah ini adalah: terjalin erat. Matriks dokumen-demi-istilah dari langkah transformasi teks berfungsi sebagai masukan data untuk sebagian besar prosedur di bagian ini. Terkadang, ketika hasilnya tidak memuaskan, peneliti dapat mempertimbangkan untuk mengubah atau memperbesar himpunan

variabel yang diturunkan dari transformasi langkah) atau memilih metode analisis lain. Biasanya berbeda kombinasi transformasi data dan teknik analisis dicoba dan diuji dan salah satu yang menghasilkan kinerja tertinggi dipilih. Sebagian besar operasi TM jatuh ke dalam salah satu dari lima jenis, yaitu, (a) pengurangan dimensi, (b) jarak dan komputasi kesamaan, (c) pengelompokan, (d) pemodelan topik, dan (e) klasifikasi. Di bawah kita diskusikan setiap teknik sebelum membahas bagaimana menilai kredibilitas dan validitas hasil text mining.

- **Dimensionality reduction**

Pengurangan dimensi. Matriks dokumen demi suku cenderung memiliki banyak variabel. Biasanya diinginkan untuk mengurangi ukuran matriks ini dengan menerapkan teknik pengurangan dimensi. Beberapa manfaat dari pengurangan dimensi adalah analisis yang lebih mudah diatur, interpretasi hasil yang lebih besar (misalnya, lebih mudah untuk menafsirkan hubungan variabel ketika jumlahnya sedikit), dan banyak lagi. representasi yang efisien. Dibandingkan dengan bekerja dengan matriks dokumen per suku awal, pengurangan dimensi juga dapat mengungkapkan dimensi laten dan menghasilkan peningkatan kinerja. Dua pendekatan umum biasanya digunakan untuk mengurangi dimensi. Salah satunya adalah membangun variabel laten baru dan yang kedua adalah menghilangkan variabel yang tidak relevan.

Dalam kasus sebelumnya, variabel baru dimodelkan sebagai kombinasi (non)linier dari variabel asli dan mungkin ditafsirkan sebagai konstruksi laten (misalnya, kata-kata tahun, pengalaman, dan diperlukan dapat digabung menjadi) mengungkapkan konsep pengalaman kerja dalam lowongan pekerjaan). Latent Semantic Analysis (LSA) biasanya digunakan untuk mendeteksi sinonim (yaitu, kata-kata berbeda yang memiliki arti yang sama) dan polisemi (yaitu, satu kata yang digunakan dalam arti yang berbeda namun terkait) di antara kata-kata. Principal Component Analysis (PCA) efektif untuk data reduksi karena mempertahankan varians data. Analisis paralel adalah strategi yang direkomendasikan untuk pilih berapa banyak dimensi yang akan dipertahankan di PCA. Kerugian dari LSA dan PCA adalah mungkin sulit untuk melampirkan makna pada dimensi yang dibangun.

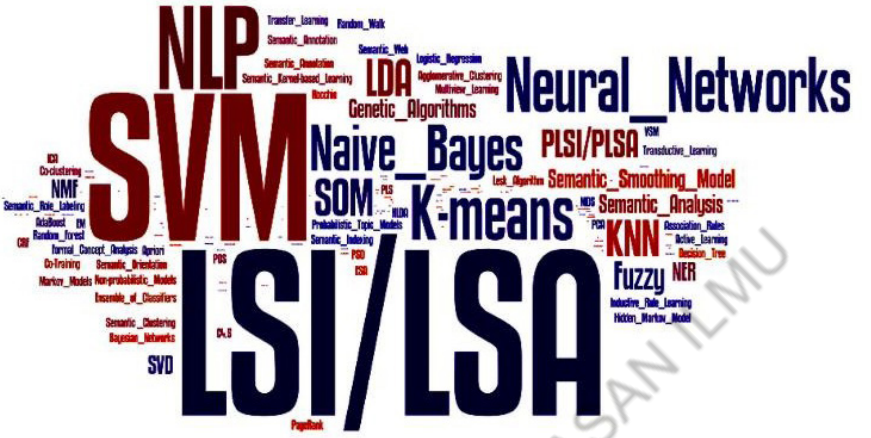
Teknik lain adalah proyeksi acak, di mana titik data diproyeksikan ke dimensi yang lebih rendah dengan tetap menjaga

jarak antara poin. Pendekatan alternatif untuk mengurangi dimensi adalah dengan menghilangkan variabel dengan menggunakan variabel metode seleksi. Berbeda dengan metode proyeksi, pemilihan variabel metode tidak membuat variabel baru melainkan memilih dari variabel yang ada dengan menghilangkan yang tidak informatif atau berlebihan (misalnya, kata-kata yang muncul di terlalu banyak dokumen mungkin tidak berguna untuk mengkategorikan dokumen). Tersedia tiga jenis metode: filter, pembungkus, dan metode tertanam. Filter menetapkan skor ke variabel dan menerapkan ambang batas skor untuk dihapus variabel yang tidak relevan. Filter populer adalah ambang batas TF- IDF, perolehan informasi, dan chi-kuadrat statistik. Pembungkus memilih subset terbaik dari variabel di hubungannya dengan metode analitis. Dalam metode yang disematkan, mencari subset variabel terbaik dicapai dengan meminimalkan fungsi tujuan yang secara bersamaan memperhitungkan model akun kinerja dan kompleksitas. Kinerja model dapat diukur misalnya dengan kesalahan prediksi (dalam kasus klasifikasi) dan kompleksitas diukur dengan jumlah variabel dalam model. Subset terkecil dari variabel yang menghasilkan kesalahan prediksi terendah adalah subset pilihan.

- **Komputasi jarak dan kesamaan (*Distance and similarity computing)***

Menilai kesamaan dua atau lebih dokumen adalah kegiatan utama dalam banyak aplikasi seperti dalam pengambilan dokumen (misalnya, pencocokan dokumen), dan rekomendasi sistem (misalnya, untuk menemukan produk serupa berdasarkan deskripsi atau ulasan produk). Banyak sekali langkah-langkah yang beroperasi pada representasi vektor dapat digunakan untuk menilai jarak atau kesamaan. Sebuah contoh yang terakhir adalah ukuran kosinus yang digunakan secara luas dalam pencarian informasi. Nilai untuk ukuran ini berkisar dari -1 (dua vektor menunjuk berlawanan arah) ke 1 (dua vektor menunjuk ke arah yang sama); 0 berarti kedua vektor tersebut ortogonal atau tegak lurus (atau tidak berkorelasi). Ukuran ini menilai kesamaan dua dokumen berdasarkan: frekuensi istilah yang mereka bagikan, yang diambil untuk menunjukkan kesamaan konten. Ukuran ini telah diterapkan untuk pencocokan dokumen dan mendeteksi semantik kesamaan. Dari berbagai ukuran jarak, ukuran jarak Euclidean dan Hamming yang paling umum digunakan. Tidak seperti ukuran kesamaan, nilai yang lebih tinggi untuk ukuran jarak menyiratkan ketidakmiripan. Juga ukuran jarak

harus memenuhi sifat-sifat tertentu seperti nonnegatif dan segitiga ketidaksamaan. Dalam kebanyakan kasus, ukuran kesamaan dapat diubah menjadi ukuran jarak dan sebaliknya.



Gambar 12.2 Metode dan Algoritma dalam Operasi Text Mining (Sinoara et al., 2017)

12.6 Studi Kasus

Analisis Data Teks dari Twitter tentang Kenaikan BBM dengan R Studio

Langkah pertama adalah mendapatkan data teks yang akan diolah. Data teks dalam analisis sentimen merupakan persepsi orang terhadap suatu kejadian maupun orang ataupun produk. Analisis dapat digunakan dibidang teknik industri untuk analisis citra merk untuk perumusan strategi pemasaran selanjutnya.

```
#Crawling data dari twitter
install.packages("rtweet")
install.packages("tidyverse") # Data Science Tool
install.packages("tidytext") # Untuk Text Mining dan Preprocessing
install.packages("rtweet") # Untuk akses ke Twitter API
install.packages("wordcloud2") # Untuk membuat wordcloud
```

Syarat untuk mengambil data tweet adalah memiliki akun twitter, lalu dapat ke twitter developer untuk mengambil API dengan sebelumnya mengisi beberapa form pertanyaan tentang tujuan penggunaan data.

```
library(rtweet)
## Warning: package 'rtweet' was built under R ver-
sion 4.1.3
token <- create_token(
   consumer_key = "EKO******************",
   consumer_secret =  BMb**********************
   **********************o",
   access_token = "23751**********************
   ******************",
    a c c e s s _ s e c r e t     =     " x T g -
   j************************************4y")
## Warning: `create_token()` was deprecated in
rtweet 1.0.0.
## See vignette('auth') for details
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to
see where this warning was generated.
## Saving auth to 'C:\Users\anik nur habyba\
AppData\Roaming/R/config/R/rtweet/
## create_token.rds'
trending_indo <- get_trends("indonesia")
trending_indo
```

trend	url
<chr>	<chr>
#TimnasDay	h t t p : / / t w i t t e r . c o m / search?q=%23TimnasDay
#TimnasDay	h t t p : / / t w i t t e r . c o m / search?q=%23TimnasDay
#SelamanyaGaruda	h t t p : / / t w i t t e r . c o m / search?q=%23SelamanyaGaruda
#AduBalap	h t t p : / / t w i t t e r . c o m / search?q=%23AduBalap
#AllegriOut	h t t p : / / t w i t t e r . c o m / search?q=%23AllegriOut
Vietnam	http://twitter.com/search?q=Vietnam
BRAND NEW MOOD 2	h t t p : / / t w i t t e r . c o m / search?q=%22BRAND+NEW+MOOD+2%22

trend	url
<chr>	<chr>
Armuh vs Marshel	http://twitter.com/search?q=ArmuhvsMarshel
Kenduri Swarnabhumi	http://twitter.com/search?q=%22Kenduri+Swarnabhumi%22
Juve	http://twitter.com/search?q=Juve

Next

12345

Previous

1-10 of 50 rows | 1-2 of 9 columns

```
library (dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
# Dari trending_indo
trending_indo %>%
# Lihat struktur data keseluruhan
glimpse()
## Rows: 50
## Columns: 9
## $ trend            <chr> "#TimnasDay", "#TimnasDay",
   "#SelamanyaGaruda", "#Adu~
## $ url                  <chr> "http://twitter.com/
   search?q=%23TimnasDay", "http://t~
## $ promoted_content <lgl> NA, NA, NA, NA, NA, NA, NA,
   NA, NA, NA, NA, NA, NA, N~
## $ query          <chr> "%23TimnasDay", "%23TimnasDay",
   "%23SelamanyaGaruda",~
## $ tweet_volume      <int> NA, NA, NA, NA, 14904,
   319969, 42647, NA, NA, 29508, ~
```

```
## $ place            <chr> "Indonesia", "Indonesia",
   "Indonesia", "Indonesia", "~
## $ woeid          <int> 23424846, 23424846, 23424846,
   23424846, 23424846, 234~
## $ as_of           <dttm> 2022-09-18 15:17:06, 2022-
   09-18 15:17:06, 2022-09-18~
## $ created_at      <dttm> 2022-09-17 06:26:32, 2022-
   09-17 06:26:32, 2022-09-17~
```

Penjelasan sebagian isi kolom:

Kolom trend berisi tren yang saat ini ada/muncul di daerah Indonesia

Kolom url berisi alamat url terkait trending topics yang dilihat

Kolom tweet_volume berisis jumlah banyaknya topik/hashtag tersebut digunakan oleh pengguna twitter

Kolom place dan woeid adalah negara dan id negara dimana trending topics diperoleh

Kolom created_at adalah waktu kapan trending topics tersebut muncul

Mengecek nama kolom-kolom variabel kembali yang akan diseleksi

```
colnames(trending_indo)
## [1] "trend"    "url"   "promoted_content"   "query"
## [5] "tweet_volume"  "place"    "woeid"      "as_of"
## [9] "created_at"
```

Pilih data kolom yang akan diambil

```
#membingkai data
trending_indo<-as.data.frame(trending_indo)
#memilih data untuk kolom trend dan tweet_volume saja
trending_indo %>% select(., trend, tweet_volume)
```

trend	tweet_volume
<chr>	<int>
Robi Darwis	NA
Marselino	NA
Udinese	24067
Iwan Bule	NA
Di Maria	14350

trend	tweet_volume
<chr>	<int>
Shin Tae Yong	NA
Nguyen	11701
Besok Senin	11657
Arsenal	350918
Koplo Superstar	NA

Next

12345

Previous

11-20 of 50 rows

Mengeliminasi nilai NA dengan menggunakan fungsi filter()

```
# Dari trending_indo
trending_indo %>%
# pilih kolom trend dan kolom tweet_volume
 select(trend, tweet_volume) %>%
# filter kolom tweet_volume dimana nilainya tidak boleh 'NA'
filter(tweet_volume != "NA")
```

trend	tweet_volume
<chr>	<int>
#AllegriOut	14904
Vietnam	319969
BRAND NEW MOOD 2	42647
Juve	29508
Udinese	24067
Di Maria	14350
Nguyen	11701
Besok Senin	11657
Arsenal	350918
Fabio	47889

Melihat tren mana yang memiliki volume terbesar dengan menggunakan fungsi arrange() dan menggunakan fungsi desc() untuk mengurutkan data berdasarkan kolom tertentu

```
# Dari trending_indo
trending_indo %>%
  # pilih kolom trend dan kolom tweet_volume
 select(trend, tweet_volume) %>%
  # filter kolom tweet_volume dimana nilainya tidak
boleh 'NA'
 filter(tweet_volume != "NA") %>%
  # disusun berdasarkan tweet_volume secara
descending
 arrange(desc(tweet_volume))
```

trend	tweet_volume
<chr>	<int>
Arsenal	350918
Vietnam	319969
GTA 6	188982
Nico	167899
Brentford	71711
Fabio	47889
Vieira	47147
Ethan Nwaneri	47078
BRAND NEW MOOD 2	42647
Roma	37300

Membuat wordcloud dari tren twitter install packages wordcloud2

```
library(wordcloud2)
## Warning: package 'wordcloud2' was built under R
version 4.1.3
# Dari trending_indo
trending_indo %>%
  # pilih kolom trend dan kolom tweet_volume
  select(trend, tweet_volume) %>%
  # filter kolom tweet_volume dimana nilainya tidak
boleh 'NA'
  filter(tweet_volume != "NA") %>%
   # disusun berdasarkan tweet_volume secara
```

```
descending
  arrange(desc(tweet_volume)) %>%
  # buat wordcloud dengan ukuran huruf 2
  wordcloud2(size = 2)
```

Gambar 12.3 Worcloud Trending Topic Twitter

###Analisa Tweet Berdasarkan Hashtag, Mention atau Keyword
Kita bisa melakukan pencarian untuk topik tertentu berdasarkan hashtag, mention maupun kata kunci apapun menggunakan fungsi search_tweet. Sebagai contoh saya mencari hashtag #pertamina

 # Cari tweet dengan hashtag #kenaikanBBM sebanyak 100 lalu masukkan ke variabel 'hasil_pencarian'

hasil_pencarian <- search_tweets(q = "#kenaikanBBM", n = 1000)

hasil_pencarian

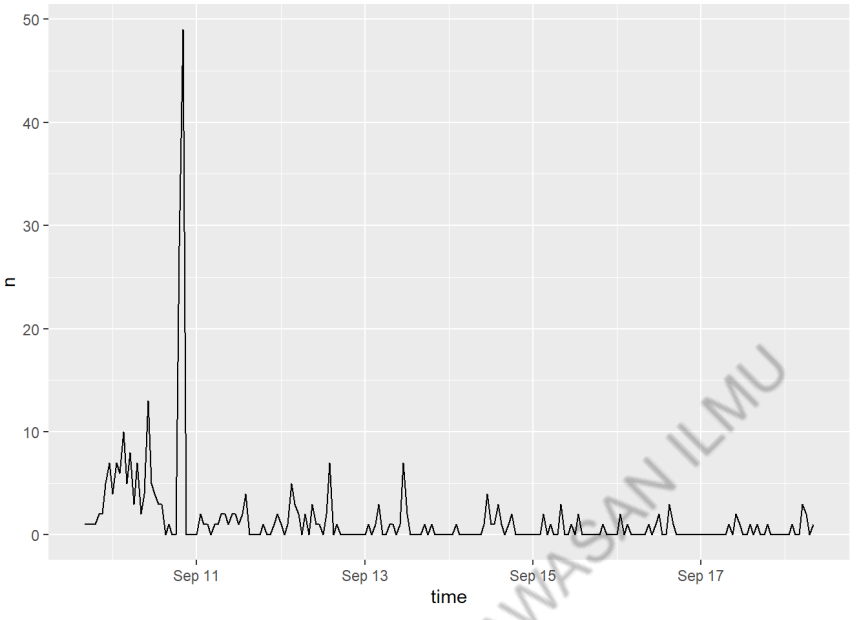
created_at	id	id_str
<dttm>	<dbl>	<chr>
2022-09-18 15:06:05	1.571410e+18	1571410214452342786
2022-09-18 13:00:25	1.571379e+18	1571378591279247360
2022-09-18 13:00:08	1.571379e+18	1571378517849559041
2022-09-18 12:59:38	1.571378e+18	1571378395380060161
2022-09-18 12:59:27	1.571378e+18	1571378347929915392
2022-09-18 12:56:39	1.571378e+18	1571377641705574400
2022-09-18 09:49:42	1.571331e+18	1571330595686612992

created_at	id	id_str
<dttm>	<dbl>	<chr>
2022-09-18 02:15:55	1.571216e+18	1571216396255916033
2022-09-17 23:29:40	1.571175e+18	1571174558903971846
2022-09-17 21:00:22	1.571137e+18	1571136985351462914

Mengecek jenis data

```
# Dari variabel hasil_pencarian
hasil_pencarian %>%
  # perlihatkan struktur data
  glimpse()
## Rows: 306
## Columns: 43
## $ created_at                    <dttm> 2022-09-
18 15:06:05, 2022-09-18 13:00:2~
## $ id                          <dbl> 1.571410e+18,
1.571379e+18, 1.571379e+18~
## $ id_str                                  <chr>
"1571410214452342786", "1571378591279247~
## $ full_text                 <chr> "BBM NAIK??\
nPro kontra kenaikan harga B~
## $ truncated                   <lgl> FALSE,
FALSE, FALSE, FALSE, FALSE, FALSE~
## $ display_text_range          <dbl> 206, 139,
140, 140, 139, 273, 99, 162, 1~
## $ entities                    <list> [[<data.
frame[3 x 2]>], [<data.frame[1 ~
## $ metadata                    <list> [<data.
frame[1 x 2]>], [<data.frame[1 x~
## $ source                          <chr> "<a
href=\"http://twitter.com/download/a~
## $ in_reply_to_status_id        <dbl> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ in_reply_to_status_id_str    <chr> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ in_reply_to_user_id          <dbl> NA, NA,
```

```
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ in_reply_to_user_id_str       <chr> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ in_reply_to_screen_name       <chr> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ geo                           <list> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA,~
## $ coordinates                   <list> [<data.
frame[1 x 3]>], [<data.frame[1 x~
## $ place                         <list> [<data.
frame[1 x 3]>], [<data.frame[1 x~
## $ contributors                   <lgl> NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ is_quote_status                <lgl> FALSE,
FALSE, FALSE, FALSE, FALSE, FALSE~
## $ retweet_count                 <int> 0, 2, 3,
3, 2, 2, 0, 0, 0, 0, 1, 0, 1, 0~
## $ favorite_count                <int> 0, 0, 0,
0, 0, 2, 1, 0, 1, 0, 0, 0, 1, 0~
## $ favorited                      <lgl> FALSE,
FALSE, FALSE, FALSE, FALSE, FALSE~
## $ retweeted                      <lgl> FALSE,
FALSE, FALSE, FALSE, FALSE, FALSE~
## $ possibly_sensitive             <list> FALSE,
NA, NA, NA, NA, FALSE, FALSE, FA~
# Dari hasil pencarian
hasil_pencarian %>%
  # buatkan grafik time-series frekuensi kata kunci
tersebut
  # dipakai dalam tweet per menit
  ts_plot(by = "hours")
## Warning in eval_tidy(x[[2]], data, env):
restarting interrupted promise
## evaluation
```

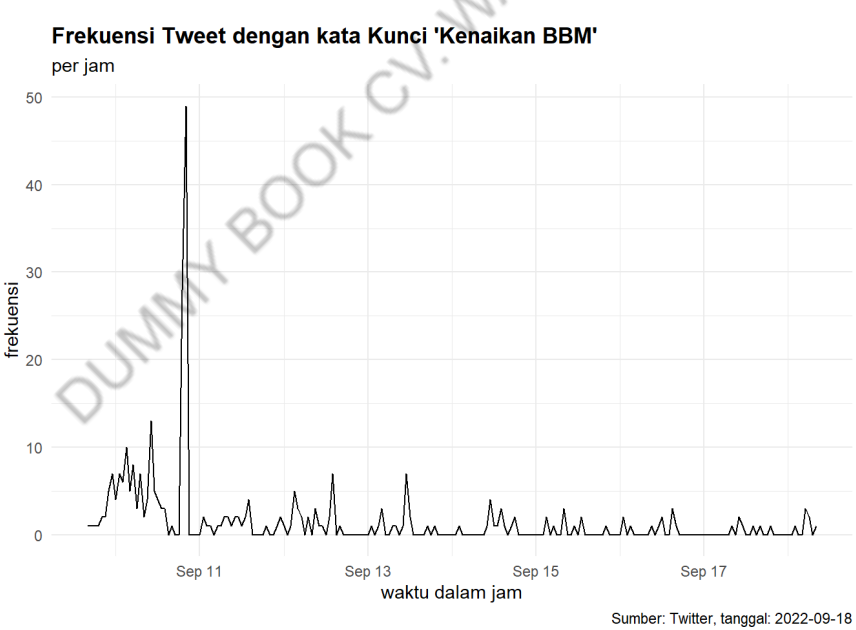


Gambar 12.4 Grafik Tweet tentang Kenaikan BBM (1)

Membuat grafik lebih menarik

```
library(ggplot2)
# Dari hasil pencarian
hasil_pencarian %>%
  # buatkan grafik time-series frekuensi
  # kata kunci tersebut dipakai dalam
  # tweet per menit
  ts_plot(by = "hours") +
  # Menggunakan theme_minimal
  theme_minimal() +
  # Memberikan judul plot dengan
  # elemen teks tebal (bold)
  theme(plot.title = element_text(face = "bold")) +
  # Menambahkan beberapa elemen
  labs(
```

```
    # Berikan label untuk x
    x = "waktu dalam jam",
    # Berikan label untuk y
    y = "frekuensi",
    # Berikan judul
     title = "Frekuensi Tweet dengan kata Kunci
'Kenaikan BBM'",
    # Memberi sub-judul
    subtitle = "per jam",
    # Memberi caption
    caption = paste0("Sumber: Twitter, tanggal: ",
Sys.Date())
  )
## Warning in eval_tidy(x[[2]], data, env):
restarting interrupted promise
## evaluation
```



Gambar 12.5 Grafik Tweet tentang Kenaikan BBM (2)

Melihat semua tweet tersebut dengan melihat variabel text:

```
hasil_pencarian<-as.data.frame(hasil_pencarian)
# Dari hasil_pencarian
hasil_pencarian %>%
# Pilih kolom teks
select(text)
```

text
<chr>
BBM NAIK??\nPro kontra kenaikan harga BBM tergantung sudut pandang rakyat terhadap pemerintah??....\n\nYuk bisa cek dan baca selengkapnya di \nhttps://t.co/L5jgjrl4Q6\n\n#kenaikanbbm #bbmnaik #kebijakanpemerintah https://t.co/MYXYyp4Avl
RT @BungaHariyono: Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai 3 September 2022…
RT @KimHaura1: Pada 3 September 2022, pemerintah akhirnya mengumumkan bahwa harga Bahan Bakar Minyak (BBM) bersubsidi jenis Pertalite dan S…
RT @KimHaura1: Pada 3 September 2022, pemerintah akhirnya mengumumkan bahwa harga Bahan Bakar Minyak (BBM) bersubsidi jenis Pertalite dan S…
RT @BungaHariyono: Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai 3 September 2022…
Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai 3 September 2022\n\n"2022 Harga BBM Naik"\nhttps://t.co/3bgsHNH9Wy\n\nKreator : Bunga Prameswari H._220503110075\n\n#pbs22b1 \n#tugaspancasila \n#kenaikanbbm \n#kebijakanpemerintah https://t.co/i926D8mw4E
#opini #kenaikanbbm #setujubbmnaik #pancasilaauinma22 #pbs22a1\n@edhenkbaru \nhttps://t.co/YuoJGEtRk2 https://t.co/HYnME4i9hG
https://t.co/CUbzIlvf27\n#parboaboa #Haiti #demonstrasi #kenaikanBBM #PBB #krisisekonomi #internasional #KabarInternasional #InfoInternasional #beritainternasional

text
<chr>
Konten ini telah tayang di https://t.co/nlBBcz6Qi7 dengan judul "Kenaikan BBM Membuat Masyarakat Menjerit", Klik untuk baca:\nhttps://t.co/8QoGegLHsa\n\n#pancasilaauinma22\n#pb-s22a1\n#kenaikanbbm https://t.co/Mb45mthwPW
https://t.co/spUcOB69Xy\n#parboaboa #padangsidempuan #bulog #pasarmurah #kenaikanBBM #Daerah #beritadaerah #sumut #sumutupdate #sumutpaten #SumutBermartabat #beritasumut

Next

123456

...

31

Previous

1-10 of 306 rows

Jika dilihat, terdapat beberapa teks yang mengandung hashtag (#), link, mention (@) dan kadangkala terdapat teks yang memuat emoticon. Hal ini bisa dibersihkan dengan menggunakan fungsi mutate, gsub() dan plain_tweets():

```
# Dari hasil_pencarian
hasil_pencarian %>%
  # Pilih kolom text
  select(text) %>%
  # Ubah elemen pada kolom text dengan mengganti
  # semua link dengan karakter kosong
  mutate(text = gsub(pattern = "http\\S+",
                     replacement = "",
                     x = text)) %>%
  mutate(text = gsub(pattern = "#",
                     replacement = "",
                     x = text)) %>%
  mutate(text = gsub(pattern = "@",
```

```
                    replacement = "",
                    x = text)) %>%
  mutate(text = gsub(pattern = "%",
                    replacement = "",
                    x = text)) %>%
  mutate(text = gsub(pattern = "[0-9]*",
                    replacement = "",
                    x = text)) %>%
  mutate(text = gsub(pattern = "/",
                    replacement = "",
                    x = text)) %>%
  mutate(text = gsub(pattern = "'",
                    replacement = "",
                    x = text)) %>%
  mutate(text = gsub(pattern = ":",
                    replacement = "",
                    x = text)) %>%

# lalu simpan ke dalam variabel text_cleaned
  plain_tweets() -> text_cleaned
## Warning: `plain_tweets()` was deprecated in
rtweet 1.0.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to
see where this warning was generated.
text_cleaned
```

text
<chr>
BBM NAIK?? Pro kontra kenaikan harga BBM tergantung sudut pandang rakyat terhadap pemerintah??.... Yuk bisa cek dan baca selengkapnya di kenaikanbbm bbmnaik kebijakanpemerintah

text
<chr>
RT BungaHariyono Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai September
RT KimHaura Pada September , pemerintah akhirnya mengumumkan bahwa harga Bahan Bakar Minyak (BBM) bersubsidi jenis Pertalite dan S
RT KimHaura Pada September , pemerintah akhirnya mengumumkan bahwa harga Bahan Bakar Minyak (BBM) bersubsidi jenis Pertalite dan S
RT BungaHariyono Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai September
Presiden Jokowidodo resmi menaikkan harga bahan bakar (BBM) bersubsidi jenis pertalite dan solar mulai September " Harga BBM Naik" Kreator Bunga Prameswari H._ pbsb tugaspancasila kenaikanbbm kebijakanpemerintah
opini kenaikanbbm setujubbmnaik pancasilaauinma pbsa edhenkbaru
parboaboa Haiti demonstrasi kenaikanBBM PBB krisisekonomi internasional KabarInternasional InfoInternasional beritainternasional
Konten ini telah tayang di dengan judul "Kenaikan BBM Membuat Masyarakat Menjerit", Klik untuk baca pancasilaauinma pbsa kenaikanbbm
parboaboa padangsidempuan bulog pasarmurah kenaikanBBM Daerah beritadaerah sumut sumutupdate sumutpaten SumutBermartabat beritasumut

Next
123456
...
31
Previous
1-10 of 306 rows

Tweet yang tersaring merupakan objek text yang merupakan character vector. Selanjutnya, kita modifikasi vektor text tersebut sehingga tiap elemen di dalamnya merupakan kata tunggal.

```
text <- unlist(strsplit(text_cleaned$text, "\\
W+"))
text
##    [1] "BBM"                "NAIK"
##    [3] "Pro"                "kontra"
##    [5] "kenaikan"           "harga"
##    [7] "BBM"                "tergantung"
##    [9] "sudut"              "pandang"
##   [11] "rakyat"             "terhadap"
##   [13] "pemerintah"         "Yuk"
##   [15] "bisa"               "cek"
##   [17] "dan"                "baca"
## [5099] "umum"               "yang"
## [5101] "pengelolaan"        "RT"
## [5103] "muslimahnewscom"    "KenaikanBBM"
## [5105] "BBM"                "adalah"
## [5107] "kebutuhan"          "vital"
## [5109] "masyarakat"         "Dalam"
## [5111] "Islam"              "BBM"
## [5113] "termasuk"           "dalam"
## [5115] "kepemilikan"        "umum"
## [5117] "yang"               "pengelolaan"
```

R bersifat case-sensitive sehingga dalam pemrosesan data teks, akan lebih baik jika kita ubah text menjadi huruf kecil (lower-case) atau huruf besar (upper-case).

```
text <- tolower(text)
text
##    [1] "bbm"                "naik"
##    [3] "pro"                "kontra"
##    [5] "kenaikan"           "harga"
##    [7] "bbm"                "tergantung"
```

```
##    [9] "sudut"                 "pandang"
##   [11] "rakyat"                "terhadap"
##   [13] "pemerintah"            "yuk"
##   [15] "bisa"                  "cek"
##   [17] "dan"                   "baca"
##   [19] "selengkapnya"          "di"
##   [21] "kenaikanbbm"           "bbmnaik"
##   [23] "kebijakanpemerintah"   "rt"
##   [25] "bungahariyono"         "presiden"
## [5109] "masyarakat"            "dalam"
## [5111] "islam"                 "bbm"
## [5113] "termasuk"              "dalam"
## [5115] "kepemilikan"           "umum"
## [5117] "yang"                  "pengelolaan"
```

Terakhir, hitung frekuensi setiap kata muncul dalam naskah pidato.

```
text <- data.frame(table(text))
text
```

text	**Freq**
<fct>	<int>
	5
_	1
a	1
aceppurnama	1
ada	12
adab	1
adalah	65
adian	2
agar	1
ahy	4

Next

123456

...

98

Previous

1-10 of 978 rows

Jika diperhatikan, terdapat kata-kata seperti kata 'ini', 'di', 'dari', 'itu' dan kata-kata lainnya yang tidak memiliki makna atau memuat topik tertentu di dalam sebuah teks, kata-kata ini biasa disebut sebagai stopwords. Stopwords dapat dihilangkan dengan mudah selama kita memiliki korpus (corpus) adalah koleksi teks yang besar dan terstruktur. Biasanya digunakan untuk analisa statistik dan pengujian hipotesis terkait stopwords.

```
# Ambil stopwords dari link, beri nama kolomnya 'stopwords'
# hasilnya disimpan ke dalam variabel 'stopword_indo'
IDStopwords <- readLines("C:/BIBIB/00_BELAJAR/
Belajar R/Text Mining/IDStopwords.txt")
## Warning in readLines("C:/BIBIB/00_BELAJAR/
Belajar R/Text Mining/
## IDStopwords.txt"): incomplete final line found
on 'C:/BIBIB/00_BELAJAR/Belajar
## R/Text Mining/IDStopwords.txt'
head(IDStopwords)
## [1] "a" "b" "c" "d" "e" "f"
text2 <- text[!is.element(text$text, IDStopwords),]
text2
```

	text <fct>	Freq <int>
1		5
2	_	1
4	aceppurnama	1
6	adab	1
8	adian	2
10	ahy	4
11	aja	2
12	akalakalan	2
15	akibat	1
16	aksi	14

Next
123456
...
84
Previous
1-10 of 837 rows

Mengurutkan text berdasarkan frekuensi terbanyak

```
library(dplyr)
sorted_text2 <- text2 %>%
  as.data.frame() %>%
  arrange(desc(Freq))
sorted_text2
```

text <fct>	Freq <int>
bbm	405
kenaikan	243
kenaikanbbm	225
sejarah	154
rt	149
presiden	81
wajib	77
muslimahnewscom	74
masyarakat	71
islam	66

Next
123456
...
84
Previous
1-10 of 837 rows

Menyaring text yang frekuensi lebih besar dari 10

```
text3<- sorted_text2 %>% filter(Freq > 10)
text3
```

text <fct>	Freq <int>
bbm	405
kenaikan	243
kenaikanbbm	225
sejarah	154
rt	149
presiden	81
wajib	77
muslimahnewscom	74
masyarakat	71
islam	66

Next

1234

Previous

1-10 of 38 rows

Membuat Wordcloud untuk kata tweet yang lebih besar dari 10 kali muncul

```
library(wordcloud2)
wordcloud2(data = text3)
```

Gambar 12.6 Wordcloud Tweet tentang Kenaikan BBM (1)

Memberikan background

```
wordcloud2(text3, color = "random-dark", back-
groundColor = "lightgrey"
```

Gambar 12.7 Wordcloud Tweet tentang Kenaikan BBM (2)

Analisis Sentimen

Analisis sentimen dilakukan untuk melihat klasifikasi tweet teks pada emosi manusia. Kamus emosi manusia diaplikasikan menggunakan NRC Emotion Lexicon atau dikenal EmoLex. Kamus ini merupakan suatu lexicon yang berisi daftar kata dan kaitannya dengan delapan emosi dasar yaitu anger, fear, anticipation, trust, surprise, sadness, joy, dan disgust, serta dua valensi sentimen yaitu negatif dan positif. Pada R, untuk memanggil kamus sentimen NRC dengan tujuan menghitung keberadaan delapan emosi yang berbeda dan valensinya yang sesuai dalam file teks dapat dilakukan dengan menggunakan fungsi get_nrc_sentiments yang didapatkan dari syuzhet.

Analisis Sentimen Mahasiswa untuk Perbaikan Desain Afektif Ruang Kelas Jurusan Teknik Industri, Universitas Trisakti (Habyba et al., 2021)

Perbaikan desain afektif ruang kelas penting untuk diakukan guna meningkatkan kinerja pembelajaran mahasiswa. Pencapaian kinerja mahasiswa Teknik Industri Universitas Trisakti dipengaruhi oleh desain afektif ruang kelas. Tujuan penelitian ini adalah menggunakan analisis sentimen dalam klasifikasi persepsi

mahasiswa terhadap desain afektif ruang kelas. Klasifikasi sentimen mahasiswa dilakukan menggunakan Support Vector Machine (SVM). Hasil analisis kuesioner juga didapatkan hasil persepsi tentang mata kuliah yang dianggap paling sulit (statistika) data dilihat pada Gambar 13.7 di bawah.

Gambar 12.8 Wordcloud Mata Kuliah Tersulit

Selain itu juga didapatkan informasi tentang ruang kelas memiliki sentimen positif yaitu FGTSC yang dapat digunakan sebagai bahan pertimbangan dalam menentukan sampel untuk tahapan perumusan desain ruang kelas selanjutnya (Gambar 13.8).

Gambar 12.9 Wordcloud Ruang Kelas Ter-Positif

Hasil menunjukkan kesan maupun hal apa yang menjadi faktor pertimbangan mahasiswa memilih kelas FGTSC. Beberapa contoh kata kansei yang dominan yaitu “nyaman” (Gambar 13.9), hal

ini menunjukkan mahasiswa sangat peduli dengan kenyamanan sebuah ruang kelas dalam proses pembelajaran.

Gambar 12.10 Wordcloud Kata Kansei tentang Ruang Kelas

Kata kansei untuk konsep desain dikumpulkan dari persepsi mahasiswa yang mendapatkan label "positif". Elemen desain yang perlu diperbaiki seperti peralatan yang kurang nyaman digunakan, pencahayaan yang kurang, tembok yang terdapat coretan serta tempat duduk yang kurang nyaman. Hasil klasifikasi menggunakan tiga SVM tipe Kernel linear, radial dan polynomial diperoleh linear memiliki nilai akurasi terbaik (76%). Hasil ini menunjukkan bahwa klasifikasi sentiment mahasiswa memiliki hasil maksimal dengan SVM tipe Kernel linear (dot). Metode ini akan digunakan dalam melakukan klasifikasi sentimen mahasiswa terhadap hasil perbaikan desain ruang kelas.

12.7 Latihan Soal

1. Jelaskan 7 Teknik text mining
2. Jelaskan contoh penggunaan Text Mining pada perusahaan

DAFTAR PUSTAKA

Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, 2013. "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, June 2013.

Agrawal A. dan Gupta. H. 2013. *Global K-Means* (GKM) *Clustering Algorithm: A Survey.* International Journal of Computer Applications, 2013. 79. 20-24. 10.5120/13713-1472.

Agusta Y. 2007. K-means penerapan, permasalahan dan metode terkait. *Jurnal Sistem dan Informatika.* 3 : 47-60.

Alfiandi A. 2021. *Perbaikan Kualitas Produk Db-Customer Display Product (DB-CDP) Dengan Integrasi Metode Cross Industry Standard Process-Data Mining (Crisp-DM) Dan Six Sigma*. Tugas Akhir. Universitas Trisakti.

Anurag kumar and Kumar Ravi Singh. 2017. "A Study on Web Content Mining," International Journal Of Engineering And Computer Science, vol. 6, no. 1, pp. 20003-20006, January 2017.

Ayele, W. Y. 2020. *Adapting CRISP-DM for Idea Mining A Data Mining Process for Generating Ideas Using a Textual Dataset* 2020. [Online]. Available:www.ijacsa.thesai.org

Bangoria B., Mankad N., dan Pambhar V. 2013. *A Survey on Efficient Enhanced K-Means Clustering Algorithm.* International Journal for Scientific Research & Development, I(9), pp.1698-700, 2013.

Benoit FD dan Van DPD. 2009. Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*. 36 (7) :10475-10484.

Bevington P. 2003. "Data reduction and error analysis for the physical sciences," vol. Third Editon. McGraw-Hill Education.

Boyd, Danah, dan Crawford, Kate. (2011). Six Provocations for Big Data. Dipresentasikan dalam Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Oxford, England.[pdf] Diperoleh dari http://software-studies.com/cultural_analytics/Six_Provocations_for_Big_Data.pdf

Cartaya dan Juan CC. 2008. La inteligencia empresarial y el Sistema de Gestión de Calidad ISO 9001:2000. *Ciencias de la Información*. 39 (1):31-44.

Charles Duhigg. 2012. How Companies Learn Your Secrets. [ONLINE] Available at: http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html?pagewanted=all&_r=1

Damayanti E. and Kuswayati S. 2018. "Analisis Dan Implementasi Framework CRISP-DM (Cross Industry Standard Process For Data Mining) Untuk Clustering Perguruan Tinggi Swasta," Sekolah Tinggi Teknologi Bandung.

Davies and P. Benyon. 2004. *Database Systems Third Edition*. Palgrave, Macmillan, New York.

Dumbill. E. 2012. *Big Data Now Current Perspective.* O'Reilly Media.

Eaton C., Dirk D., Tom D., George L., dan Paul Z.. 2012. *Understanding Big Data.* Mc Graw Hill.

Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, Aya Soffer. 2003. "The Connectivity Sonar: Detecting Site Functionality by Structural Patterns", In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HYPERTEXT). ACM Press, New York, NY, pp.38–47, 2003.

Faustina Johnson and Kumar Santosh Gupta. 2012. "Web Content Mining Techniques: A Survey," International Journal of Computer Applications(0975–888),vol Volume47–No.11,pp.44-50,June2012.

Fitriana R, Eriyatno, Djatna T, Kusmuljono BS. 2012. Peran sistem intelijensia bisnis dalam manajemen pengelolaan pelanggan dan mutu untuk agroindustri susu skala usaha menengah. *Jurnal Teknologi Industri Pertanian.* 22 (3) : 131 – 139.

Fitriana R, Saragih J, dan Firmansyah MA. 2016. *Business Intelligence System Model Proposals to Improve the Quality of Service at PT GIA.* Proceeding of 9th Intl Seminar on Industrial Engineering and Management. Padang, Indonesia. 20 – 22 September 2017.

Fitriana R, Saragih J, dan Lutfiana N. 2017. *Model business intelligence system design of quality products by using data mining in R Bakery Company.* Proceeding of 10th Intl Seminar on Industrial Engineering and Management. Belitung, Indonesia. 7-9 September 2017. IOP Conf. Series: Materials Science and Engineering. doi:10.,1088/1757- 899X/277/1/011001.

Fitriana R. 2013. *Rancang bangun sistem intelijensia bisnis untuk agroindustri susu skala menengah di Indonesia*. [Disertasi]. Bogor : Institut Pertanian Bogor.

Fitriana, R., Saragih, J., dan Hasyati B. A. 2018. *Perancangan Model Sistem Intelijensia Bisnis untuk Menganalisis Pemasaran Produk Roti di Pabrik Roti Menggunakan Metode Data Mining dan Cube*. Jurnal Teknologi Industri Pertanian, 28(1), p.113-126, 2018.

Fitriana, R., Saragih, J., Fauziah S.D. 2020 "Quality improvement on Common Rail Type-1 Product using Six Sigma Method and Data Mining on Forging Line in PT. ABC." *ISIEM 12.* IOP Conf. Ser.: Mater. Sci. Eng. 847 012038.

Fuentes T dan Louis. 2010. Incorporation Of Business Intelligence Elements In The Admission And Registration Process Of A Chilean University. *INGENIARE - Revista Chilena de Ingeniería*. 18(3): 383-394.

Ghazanfari M, Jafari M, dan Rouhani S. 2011. A tool to evaluate the business intelligence of enterprise systems. *Journal Scientia Iranica, Transactions E: Industrial Engineering*. 18 (6): 1579–1590.

Golfarelli M, Mandreoli F, PenzoW, Rizzi S, Turricchia E. 2012. OLAP query reformulation in peer-to-peer data warehousing. *Information Systems*. 37(5): 393-411.

Habyba A N, Fitriana R. 2020. *Improvement of insurance agents performance using data mining in OV agency*. International Conference on Advanced Mechanical and Industrial engineering. IOP Conf. Series: Materials Science and Engineering 909.

Haenlien M, Kaplan AM, dan Beeser AJ. 2007. A Model to determine customer lifetime value in a retail banking context. *European Management Journal.* 3 (5): 221-234.

Han J dan Kamber M. 2006. *Data mining – Concept and Techniques.* Ed. ke-2, Morgan-Kauffman, San Diego.

Harahab. E.F, Fitriana R. Faturrahman C.L. 2021. *Applying Data Mining of Association Rules as Decision Making In Coffee-Shop: a Case Study*. 17th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering.

Hayashi Y. 2010. Understanding consumer heterogeneity: A business intelligence application of neural networks. *Knowledge- Based Systems.* 23(8): 856-863.

Head SC, Nielson AR, dan Au MK. 2010. *Using commercial off-theshelf business intelligence software tools to support aircraft and automated test system maintenance environments.* AUTOTESTCON IEEE Conferences. Orlando, FL, USA.13 -16 September 2010.doi 10.1109/AUTEST.2010.5613548.

Hidalgo P, Manzur E, Olavarrieta S, Farías P. 2007. Customer retention and price matching: the afps case. *Journal of Business Research.* 61 (6): 691-696.

Hidayat M.K., Fitriana R..2022. Penerapan Sistem Intelijensia Bisnis Dan K-Means Clustering Untuk Memantau Produksi Tanaman Obat. Jurnal Teknologi Industri Pertanian Vol. 32 No. 2

https://www.teknologi-bigdata.com/2013/02/mapreduce-besar-dan-powerful-tapi-tidak.html

IBM Big Data & Analytics Hub. 2014. The Four V's of Big Data. [ONLINE] Available at: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

Imelda. 2011. Business Intelligence. *Jurnal Majalah Ilmiah UNIKOM*. 11(1):111-122.

Indraputra R. A., Fitriana R. 2020. *K-Means Clustering* Data COVID-19. Jurnal Teknik Induststi.

Ithape, A. (2022, May 4). *Design in GIS Applications(Geographical Information System)*. https://bootcamp.uxdesign.cc/ux-design-in-gis-applications-geographical-information-system-a3ce8fef484

Jadav JJ dan Panchal M. 2012. Association rule mining method on OLAP Cube. *International Journal of Engineering Research and Applications*.. 2 (2):1147-1151.

James E. Pitkow and Krishna A. Bharat. 1994. "Webviz: A tool for WorldWide Web Access Log Analysis," In Proceedings of the First International WWW Conference, January 1994.

Jie H. 2010. Research on mechanism and applicable framework of e-business intelligence. (ICEE). International Conference *E-Business and E- Government*. Guangzhou, China. 7-9 May 2010. doi. 10.1109/ICEE.2010.57.

John M. Pierre, 2001. "On the Automated Classification of Web Sites, Link¨oping Electronic Articles in Computer and Information Science, Vol.6, http://www.ep.liu.se/ea/cis/2001/000/, (Accessed on July 7, 2016).

K.M. Bataineh, M. Naji, dan M. Saqer. 2011. *A Comparison Study between Various Fuzzy Clustering Algorithms*. Jordan Journal of Mechanical and Industrial Engineering, Vol 5 No. 4. Irbid.

K.M. Bataineh, M. Naji, dan M. Saqer.A. 2011. *Comparison Study between Various Fuzzy Clustering Algorithms.* Jordan Journal of Mechanical andIndustrial Engineering,Vol 5 No. 4. Irbid.

Kamel R. 2002. PUZZLE: A concept and prototype for linking business intelligence to business strategy. *The Journal of Strategic Information Systems.* 11(2): 133-152.

Kanth A, Mushtag A, dan Khan RA. 2015. *Data Mining For Marketing.* Norderstedt Germany: GRIN Verlag GmbH.

Katal,A, 2013. Big Data: Issues, Challenges, Tools and Good Practices. Big Data: Issues, Challenges, Tools and Good Practices, 1, 404.

Kleesuwan S, Mithata S, Yupapin P, Piyatamrong B. 2010. Business intelligence in thailand's higher educational resources management. *Procedia Social and Behavioral Sciences.* 2: 84-87.

Ko IS dan Abdullaev SR. 2007. A Study on the Aspects of Successful Business intelligence System Development. *International Conference on Computational Science*. Beijing, China. 27-30 May 2007 Di dalamShi Y, van Albada GD, Dongarra J.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & den Hartog, D. N. 2018. Text Mining in Organizational Research. *Organizational Research Methods*, *21*(3), 733–765. https://doi.org/10.1177/1094428117722619.

Kotler P dan Keller. 2009. *Manajemen Pemasaran.* I (13). Jakarta: Erlangga.

Lansley, G., & Cheshire, J. 2020. February 4). *An Introduction to Spatial Data Analysis and Visualisation in R*. https://data.cdrc.ac.uk/dataset/introduction-spatial-data-analysis-and-visualisation-r

Larose DT. 2006. *Data Mining Methods And Models*. New Jersey: John Wiley & Sons, Inc., Hoboken.

Li ST, Shue LY, dan Lee SF. 2008. Business intelligence approach to supporting strategy- making of ISP service management. *Journal Expert Systems with Applications: An InternationalJournal* 35(3):739- 754.doi:10.1016/j.eswa.2007.07.049.

Li, D., Wang, S., & Li, D. 2015. *Spatial Data Mining*. Springer.

Li, D., Wang, S., & Li, D. 2016. Spatial data mining: Theory and application. In *Spatial Data Mining: Theory and Application*. https://doi.org/10.1007/978-3-662-48538-5

Liu L. 2010. Supply Chain Integration through Business Intelligence; Management and ServiceScience (MASS). *International Conference on Mathematical and Statistical Science.* Cape Town South Africa. 27-29 Januari 2010. doi: 10.1109/ICMSS. 2010.5577534.

M. Bilal, P. M. L. Chan, and W. Khan. 2013. "Cooperative Network for Vehicular Communications: Game Theoretic Distribution of Reward among Contributing Vehicles," Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT), vol. 3, no. 8, pp. 11-25, Augest 2013.

M. Hung, J. Wu, J.H. Chang, dan D. Yang. 2005. *An Efficient k-Means Clustering Algorithm Using Simple Partitioning*. Journal of Information Science and Engineering. 21. 1157-1177. 2005.

Madhulatha T. 2012. An Overview on Clustering Methods. IOSR Journal of Engineering. 2. 10.9790/3021-0204719725.

Madhulatha. T. 2012. *An Overview on Clustering Methods*. IOSR Journal of Engineering, 2012. 2. 10.9790/3021-0204719725.

Maira P. 2009. Managing sustainability with the support of business intelligence: Integrating socio-environmental indicators and organisational context. *The Journal of Strategic Information Systems.* 18 (4): 178-191.

Marjanovic O. 2010. *Business Value Creation through Business Processes Management and Operational Business Intelligence Integration.*43rd Hawaii International Conference on System Science *(HICSS).* Honolulu, HI, USA. 5-8 Januari 2010. doi : 10.1109/HICSS.2010.89.

Martinez-Plumed F. *et al.* 2021. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

Ming KC dan Shih CW. 2010. The use of a hybrid fuzzy-Delphi-AHP approach to develop global business intelligence for information service firms. *Expert Systems with Applications.* 37(11): 7394–7407.

Ming M, Zehui L, dan Jinyuan C. 2008. Phase-type distribution of customer relationship with Markovian response and marketing expenditure decision on the customer lifetime value. *European Journal of Operational Research.* 187 :313– 326.

Mughal M.J.H. 2018. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. (IJACSA) International Journal of Advanced Computer Science andApplications,Vol. 9, No. 6, 2018.

Mulyana JRP. 2014. *Pentaho:Solusi Open Source untuk Membangun Data Warehouse.* Yokyakarta: Penerbit Andi.

Nailah, A.Harsono, G.P.Liansari, 2014. Usulan Perbaikan Untuk Mengurangi Jumlah Cacatpada Produk Sandal Eiger-101 Lightspeed dengan Menggunakan Metode Six Sigma, *Jurnal Online Institut Teknologi Nasional*. Vol.2, pp.257-258, 2014.

Phan DD danVogel DR. 2010. A model of customer relationship management and business intelligence systems for catalogue and online retailers. *Information & Management*. 47(2): 559-566.

Pillai J. 2011. User centric approach to itemset utility mining in Market Basket Analysis. *International Journal Computer Science & Engineering*. 3 (1): 393-400.

Pol K., Patil N., Patankar S., and Das C. 2008. *A Survey on Web Content Mining and extraction of Structured and Semistructured data,* Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.

Polyvyanyy A, Ouyang C, Barros A, van der Aalst WMP. 2017. *Process querying: Enabling business intelligence through query-based process analytics. Decision Support Systems.* 100: 41-56.

Pramudiono I.. 2007. *Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data*. Kuliah Umum IlmuKomputer. Com.

Pranit Bari and P.M. Chawan. 2013. "Web Usage Mining," Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, pp. 34- 38, June 2013.

Qwaider QW. 2012. Apply on-line analytical processing (OLAP) with data mining for clinical decision sopport. *International Journal of Managing Information Technology*. 4(1): 25-37. doi:10.5121/ijmit.2012.4103. 25.

R. Malarvizhi and K Saraswathi. 2013. "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940- 2945, Augest 2013.

Romeijn, H. 2022. *Raster and Vector Data in GIS, Weighing the advantages and disadvantages*. https://spatialvision.com.au/blog-raster-and-vector-data-in-gis/.

S. Vijayarani and A. Sakila. 2015. "Multimedia Mining Research – An Overview," International Journal of Computer Graphics & Animation (IJCGA), vol. 5, pp. 69-77, January 2015.

Santosa B. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta : Graha Ilmu.

Santosa B.,Umam, A. 2018. Data Mining dan Big Data Analytics. Yogyakarta: Media Pustaka.

Saragih J., Fitriana R., Adriyan T. 2020. *Quality Improvement for Product Body 2-1 at PT. X.* ISIEM 12. IOP Conf. Series: Materials Science and Engineering 847.

SaS. 2015. Three big benefits of big data analytics. [ONLINE] Available at: http://www.sas.com/en_sa/news/sascom/2014q3/Big-datadavenport.html. [Accessed 16 December 15].

Sinoara, R. A., Antunes, J., & Rezende, S. O. 2017. Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society*, *23*(1). https://doi.org/10.1186/S13173-017-0058-7.

Sloot PMA (ed), Computer Science – ICCS. 2007. Lecturer Note in Book Science. Beijing China :Springer.

Stanton J. 2013. "Data Science Version 3: An Introduction To," Syracuse University.

Stefanovic N dan Stefanovic D. 2009. *Supply Chain Business Intelligence: Technologies, Issues and Trends* di dalam *M. Bramer(Ed.): Artificial Intelligence an Intl Perspective*. Lecture Notes in Computer Science.5640. Berlin Heidelberg: Springer. P217-245.

Steinberg D. 2009. *Chapter 10 CART: Classification and Regression Trees*. Taylor & Francis Group, LLC.

Studer S. Bui T. B., Drescher C., Hanuschkin A., Winkler L., Peters S., Mülle K. R. 2021. "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, Apr. 2021.

Tan, Steinbach Kumar. 2006. *Introduction to Data Mining*, Pearson Addison Wiley.

The Big Data Landscape. 2014. Why Big Data Is A Must In Ecommerce. [ONLINE] Available at: http://www.bigdata-landscape.com/news/why-big-data-is-a-must-inecommerce. [Accessed 09 December 15.

Tina R. Patil and S. S. Sherekar. 2013. "16. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal Of Computer Science And Applications, vol. 6, pp. 256-261, April 2013.

Varghese, Bindiya dan Unnikrishnan, Avittathur, dan K. Jacob. 2011. *Clustering Student Data to Characterize Performance Patterns*. International Journal of Advanced Computer Science and Applications Special Issue, 2011. 10.14569/ SpecialIssue.2011.010322.

Vercellis C. 2009. *Business intelligence : data mining and optimization for decision making*. Chichester: John Wiley & Sons.

Whitten BD. 2004. *System Analysis and Design Methods*. 6th Ed. The Mc Graw-Hill Companies.

Witten I.H. dan Frank E. 2006. *Data Mining: Practical Machine Learning Tools and Techniques*. Biomedical Engineering Online – BIOMED ENG ONLINE. 5. 1-2.

Xiaoguang QI, Brian D. Davision. 2009. "Web Page Classification: Features and Algorithms", ACM comput. Survey 41, Article 12, http://doi.acm.org/10.1145/ 1459352. 1459357.

Zhang, Harry, 2004, The Optimality of Naive Bayes. FLAIRS conference.

INDEKS

PROFIL PENULIS

Rina Fitriana lulus dari Jurusan Teknik Industri FTI Universitas Trisakti di Jakarta pada tahun 1998. Pada tahun 2000 menyelesaikan Pendidikan S2 pada Program Magister Manajemen di Sekolah Tinggi Manajemen PPM di Jakarta. Pada tahun 2013 menyelesaikan Pendidikan Doktornya pada Program Studi Teknologi Industri Pertanian Institut Pertanian Bogor. Rina pertama kali berkarir sebagai Dosen pada tahun 2002 di Jurusan Teknik Industri Fakultas Teknologi Industri Universitas Trisakti. Rina dipilih menjadi Sekertaris Jurusan Teknik Industri FTI Usakti pada tahun 2014. Pada tahun 2017 sampai sekarang dipercaya menjadi Ketua Jurusan Teknik Industri FTI Universitas Trisakti. Beberapa mata kuliah yang diampu yaitu Analisis Perancangan Sistem Informasi, Pengendalian dan Penjaminan Mutu, Data Mining, Business Intelligence, Simulasi Sistem, Statistika Multivariat. Saat ini Rina aktif meneliti di bidang Rekayasa Kualitas, Sistem dan Simulasi Industri, Data Mining, Business Intelligence. Sebagai penunjang Rina memiliki Sertikasi Dosen, Certified International Business Intelligence Associate and Proffesional, Certified in Big Data Associate dan Insinyur Profesional Madya dan Asean Engineer. Rina telah menulis jurnal dan seminar baik nasional maupun internasional terindeks scopus dan juga buku ajar Pengendalian dan Penjaminan Mutu dan Analisis Perancangan Sistem informasi. Rina dapat dihubungi pada email : rinaf@trisakti.ac.id

Anik Nur Habyba lulus dari Universitas Brawijaya di Malang pada Jurusan Teknologi Industri Pertanian pada tahun 2015. Pada tahun 2016, Bibib melanjutkan pendidikannya pada program Magister Teknologi Industri Pertanian di IPB University dan lulus di tahun 2018. Bibib memulai karir sebagai dosen di Jurusan Teknik Industri di Universitas Trisakti sejak Juli 2019 sampai saat ini. Beberapa mata kuliah yang diampu yaitu Pengendalian dan Penjaminan Mutu, Statistika, Statistika Industri, Perancangan Eksperimen, Ekonomi Teknik, Psikologi Industri dan Kewirausahaan Berbasis Teknologi. Saat ini Anik aktif sebagai anggota Laboratorium Rekayasa Kualitas serta Laboratorium

Perancangan Organisasi dan Bisnis. Saat ini Bibib aktif dalam bidang penelitian yang terkait tentang rekayasa kualitas, data mining, serta perancangan produk dengan pendekatan Kansei Engineering. Bibib telah menulis jurnal dan seminar baik nasional maupun internasional terindeks scopus dan juga buku ajar Pengendalian dan Penjaminan Mutu. Bibib dapat dihubungi melalui anik@trisakti.ac.id

Elfira Febriani menyelesaikan Pendidikan S1 dari Teknologi Industri Pertanian, Fakultas Teknologi Pertanian, Institut Pertanian Bogor (IPB), S2 Magister Teknologi Industri Pertanian IPB. Elfira bergabung menjadi Dosen Teknik Industri, Universitas Trisakti pada tahun 2015 hingga saat ini. Elfira pernah menjadi Kepala Praktikum Analisis dan Perancangan Sistem Informasi dan saat ini menjadi Kepala Laboratorium Rekayasa Kualitas. Elfira telah melakukan penelitian mengenai beberapa penelitian. Elfira telah menghasilkan jurnal terindeks scopus dan beberapa prosiding internasional dan buku ajar Analisis Keputusan. Elfira dapat dihubungi di elfira.febriani@trisakti.ac.id