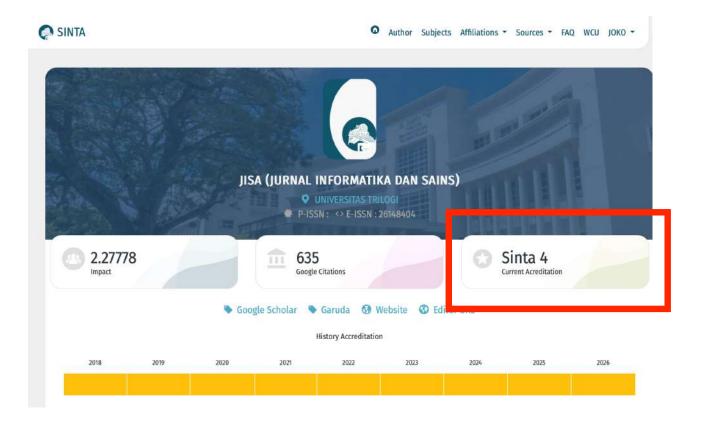
ISSN | 2614-8404





Ditterbitkon oleh :

Program Studi Teknik Informatika UNIVERSITAS TRILOGI



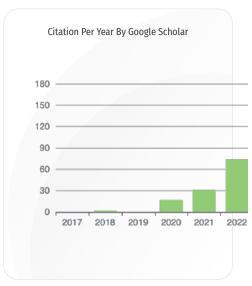








Get More with
SINTA Insight
Go to Insight



Journ	nal By Google So	cholar
	All	Since 2020
Citation	635	630
h-index	12	12
i10-index	18	18

■ 0 cited

₫ 2024

Long Short Term Memory For Comparison Between Bank Syariah Indonesia And PT Bank Artos Indonesia Shares

Authors: ZD Rizqiana, IM Akhsan, II Priyanto, AS Maharani
Sains) 7 (2), 226-232, 2024

<u>□ 2024</u> <u>□ 0 cited</u>

<u>Vehicular Ad-Hoc Networks for Intelligent Transportation System: A Brief Review of Protocols, Challenges, and Future Research</u>

Authors: KBY Bintoro IJISA (Jurnal Informatika dan Sains) 7 (2), 206-216, 2024

<u>□ 2024</u> <u>□ 1 cited</u>

<u>Design and Development of an Android-Based Merchandise Management</u> <u>Application for K-pop Consignment Services</u>

<u>□ 2024</u> <u>□ 0 cited</u>

<u>Development of the Story of Life: A Narrative and Educational Game Using the Godot Engine for Android</u>

Authors: A Kusheryanto, A Mirza, A Syaripudin, DD Hutagalung dan Sains) 7 (2), 115-124, 2024

<u>□ 2024</u> <u>□ 1 cited</u>

<u>Development of a React Native-Based Mobile E-Commerce Application to Optimize</u> <u>Online Sales for MSMEs: A Case Study of New Delisio Bakery Cake</u>

Authors: WK Suryanto, SD Sancoko JISA (Jurnal Informatika dan Sains) 7 (2), 186-196, 2024

<u>□ 2024</u> <u>□ 1 cited</u>

<u>Development of Paramadina Roomhub Application As Room Booking System Using</u> Waterfall Method

Authors: RA Maulana, MA Fatih, LA Suto, M Darwis

7 (2), 176-185, 2024

<u>□ 2024</u> <u>□ 5 cited</u>

<u>Prediction of Carbon Emissions in Indonesia Using Machine Learning: A Focus on Environmental Impact</u>

Authors: R Putra, M Sanjaya, D Utama, B Berliana JISA (Jurnal Informatika dan Sains) 7
(2), 148-152, 2024

<u>□ 2024</u> <u>□ 0 cited</u>

<u>Previous</u> 2 3 4 5 6 <u>Next</u>

Page 4 of 18 | Total Records 171



JISA(Jurnal Informatika dan Sains)

Journal of Informatics and Science

p-ISSN: 2776-3234

HOME

ABOUT

LOGIN

REGISTER

CATEGORIES

SEARCH

CURRENT

ARCHIVES

ANNOUNCEMENTS

PUBLICATION ETHICS

INDEXING

HARDCOPY

Home > Vol 8, No 1 (2025)

JISA(Jurnal Informatika dan Sains)

IISA (Jurnal Informatika dan Sains) is an electronic publication media which publishes research articles in the field of Informatics and Sciences, which encompasses software engineering, Multimedia, Networking, and soft computing. Journal published by Program Studi Teknik Informatika Universitas Trilogi and in collaboration with the Indonesian Artificial Intelligent Ecosystem (IAIE) aims to give knowledge that can be used as a reference for researchers and can be useful for society. Accredited "SINTA 4" (Vol 5,No 1, June 2022 -Vol 9 No 2, December 2026) by The Ministry of Research-Technology and Higher Education Republic of Indonesia (SK No. 230/M/KPT/2022), Article Processing Charge will be applied when the article has been declared accepted before payment is made by the author. JISA (Jurnal Informatika dan Sains) is scheduled for publication in June and December (2 issue a year). This Journal accepts research articles in these following fields:

- 1. Software Engineering: Web Development, Mobile Apps Development, Database Management System
- 2. Multimedia: Augmented Reality, Virtual Reality, Game Development
- 3. Networking: Cloud Computing, Internet of Things, Wireless Sensor Network, Mobile Computing
- 4. Soft Computing: Data Mining, Data Warehouse, Data Science, Artificial Intelligence, Decision Support System



Iournal Profile

Last update: 10 March 2023 Number of documents: 91 Total citations (all years) = **109**

h-Index: 5, i10-Index: 2 Google Scholar URL: Click Here

Submit an Article

Article Processing Charge

Focus and Scope

Author Guidelines

Article Templates

Peer Review Process

Publication Frequency

Open Access Policy

Authors are invited to submit articles that have not been published previously and should follow the Author Guideline [here]

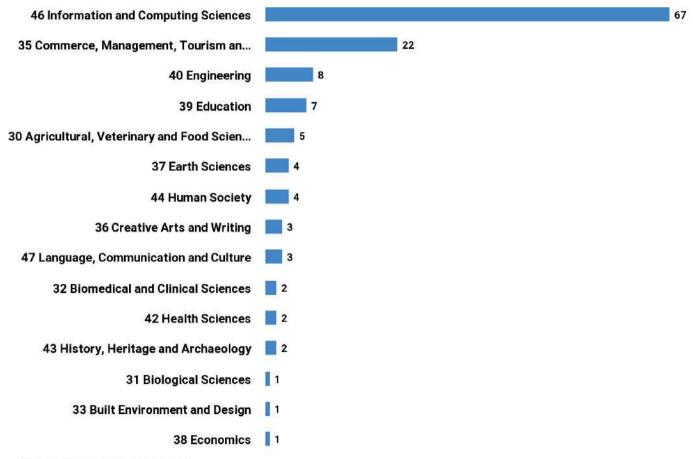
[Make a Submission]

[Download Author Guideline] [Download Article Template]

SCOPUS Citation Analysis [Click Here]

Research Field from Dimensions [Click Here]

number of publications in each research category. (Criteria: see below)



Source: https://app.dimensions.ai Exported: January 05, 2023

Criteria: 'JISA(Jurnal Informatika dan Sains)' in full data; Source Title is JISA(Jurnal Informatika dan Sains).

Copyright Agreement

Editorial Team

Peer-Reviewers

Collaboration

Statistics

Indexing

Scopus Citation Analysis











PKP INDEX

© 2023 Digital Science and Research Solutions Inc. All rights reserved. Non-commercial redistribution / external re-use of this work is permitted subject to appropriate acknowledgement. This work is sourced from Dimensions® at www.dimensions.ai.

Announcements

Article Processing Charges

From Vol 5 no 1, June 2022 Edition. This journal charges the following author fees.

Article Submission: 0.00 (IDR)

Authors are not required to pay an Article Submission Fee as part of the submission process to contribute to review costs.

Article Processing Charges: 400.000 (IDR) for Foreign Authors 50 USD

This fee includes peer-reviewing, editing, and publishing and is charge after the article is accepted.

Fast Track Review (Optional): 1.000.000 (IDR) for Indonesian Author, 100 USD for foreign Author

Posted: 2022-01-31

Publication Timeline June & December

Article Submission Deadline (June Issue): 15 May

Article Submission Deadline (December Issue): 15 November

Posted: 2021-06-08 More...

More Announcements...

Vol 8, No 1 (2025): JISA(Jurnal Informatika dan Sains)

JISA(Jurnal Informatika dan Sains) Volume 8 Issue 1 Year 2025 (June 2025) has been officially published on June 2025. This issue









In Collaboration With:





Plagiarism Tool



Grammar Tool



Vol 7, No 2 (2024) 15/10/25, 21.08



JISA(Jurnal Informatika dan Sains)

Journal of Informatics and Science

p-ISSN: 2776-3234

номе

ABOUT

USER HOME

CATEGORIES SEARCH CURRENT

ARCHIVES ANNOUNCEMENTS PUBLICATION ETHICS

INDEXING

HARDCOPY

Home > Archives > Vol 7, No 2 (2024)

Vol 7, No 2 (2024)

JISA(Jurnal Informatika dan Sains)

DOI: https://doi.org/10.31326/jisa.v7i2

JISA(Jurnal Informatika dan Sains) Volume 7 Issue 2 Year 2024 (December 2024) has been officially published on December 2024. This issue contains 15 articles from various autrhor countries which are Indonesia and Iraq.



Full Issue

View or download the full issue PDF

Table of Contents

Articles

Application of CNN in the Classification of Chili Varieties for Agricultural Efficiency Febrian Trio Pamungkas, Irsyad Zainal Muttaqin	PDF 111-114
Development of the Story of Life: A Narrative and Educational Game Using the Godot Engi for Android Andhika Kusheryanto, Anis Mirza, Ari Syaripudin, Deanna Durbin Hutagalung	ne PDF 115-124
Decision Tree for Determining Hospital Treatment for Covid-19 Patients Based on Hemato Parameters Using the C5.0 Algorithm Joko Riyono, Christina Eni Pujiastuti, Supriyadi Supriyadi, Dody Prayitno, Aina Latifa Riyana Putri	125-132
Cloud Computing Technology in Supporting the Implementation of One Data Indonesia Dara Sawitri	133-139
Smart-Working and Hot Desking Application Development using Agile and Extreme Programming Method at Xtra Cowork Daeng Ahmad Nurdin, Dodik Firmansah, Akrom Hafifi, Muhammad Darwis	PDF 140-147
Prediction of Carbon Emissions in Indonesia Using Machine Learning: A Focus on Environmental Impact Rizaldi Putra, Memet Sanjaya, Deni Utama, Berliana Berliana	PDF 148-152
Design and Development of an Android-Based Merchandise Management Application for I pop Consignment Services Nilam Putri Cahyani, Sutarman Sutarman	K- PDF 153-161
Development of Mobile GIS Based Digital Map Location Marking Application for Navigation Purposes Alif Ghifari, Sutarman Sutarman	n PDF 162-175
Development of Paramadina Roomhub Application As Room Booking System Using Water Method	fall PDF 176-185

Journal Profile

Last update: 10 March 2023 Number of documents: 91 Total citations (all years) = 109 h-Index: 5, i10-Index: 2 Google Scholar URL: Click Here

Submit an Article

Article Processing Charge

Focus and Scope

Author Guidelines

Article Templates

Peer Review Process

Publication Frequency

Open Access Policy

Copyright Agreement

Editorial Team

Peer-Reviewers

Collaboration Statistics

Indexing

Scopus Citation Analysis









Vol 7, No 2 (2024) 15/10/25, 21.08

Reza Arif Maulana, Muhamad Adillah Fatih, Lintang Arbi Suto, Muhammad Darwis	
Development of a React Native-Based Mobile E-Commerce Application to Optimize Online Sales for MSMEs: A Case Study of New Delisio Bakery Cake William Kessler Suryanto, Sulistyo Dwi Sancoko	PDF 186-196
Donation Raising Application Using Rapid Application Development (RAD) Method Based on Mobile Application Aditya Bagas Nugraha, Tri Widodo	PDF 197-205
Vehicular Ad-Hoc Networks for Intelligent Transportation System: A Brief Review of Protocols, Challenges, and Future Research Ketut Bayu Yogha Bintoro	PDF 206-216
Development of a Cashier Business Transaction System using the Android Based Agile Method David Bagus Junanda, Rodhiyah Mardhiyyah	PDF 217-225
Long Short Term Memory For Comparison Between Bank Syariah Indonesia And PT Bank Artos Indonesia Shares Zulfanita Dien Rizqiana, Izzat Muhammad Akhsan, Intan Indrasara Priyanto, Aninda Sabila Maharani	PDF 226-232
Bio-Inspired Algorithms in Healthcare Firdaws Rizgar Tato, Ibrahim Mahmood Ibrahim	PDF 233-239

JOURNAL IDENTITY

Journal Name: JISA (Jurnal Informatika dan Sains)

e-ISSN: 2614-8404, p-ISSN: 2776-3234

Publisher: Program Studi Teknik Informatika Universitas Trilogi

Publication Schedule: June and December

Language: English

APC: The Journal Charges Fees for Publishing

Indexing: EBSCO , DOAJ, Google Scholar, Arsip Relawan Jurnal Indonesia, Directory of Research Journals Indexing, Index

Copernicus International, PKP Index, Science and Technology Index (SINTA, S4) , Garuda Index

OAI address: http://trilogi.ac.id/journal/ks/index.php/JISA/oai

Contact: jisa@trilogi.ac.id

Sponsored by: DOI – Digital Object Identifier Crossref, Universitas Trilogi

In Collaboration With: Indonesian Artificial Intelligent Ecosystem(IAIE), Relawan Jurnal Indonesia, Jurnal Teknologi dan Sistem Komputer (JTSiskom)





Editorial Board 15/10/25, 20.37



JISA(Jurnal Informatika dan Sains)

Journal of Informatics and Science

REGISTER

p-ISSN: 2776-3234

LOGIN

CATEGORIES SEARCH CURRENT ARCHIVES ANNOUNCEMENTS

PUBLICATION ETHICS

INDEXING

HARDCOPY

Home > Editorial Board

ABOUT

Editorial Board

Editor in Chief

номе

Assoc.Prof.Budi Arifitama.,Ph.D (Universitas Trilogi, Indonesia | SCOPUS H-Index: 2, Google Scholar H-Index: 6)

Editorial Board

- 1. Assoc.Prof. Ir. Yaddarabullah, M.Kom, IPM,ASEAN.Eng (Universitas Trilogi, Indonesia | SCOPUS H-Index: 1, Google Scholar H-Index: 5)
- 2. Dr. Ketut Bayu Yogha Bintoro (Universitas Trilogi, Indonesia | SCOPUS H-Index: 1| Google Scholar)
- 3. Ninuk Wiliani, Ph.D (Universitas Pancasila, Indonesia | SCOPUS H-Index : 1, Google Scholar H-Index : 7)
- 4. Maya Cendana, S.T., M.Cs (Universitas Bunda Mulia, Indonesia | SCOPUS H-Index :1, Google Scholar H-Index :3)
- 5. Dwi Pebrianti, Ph.D (International Islamic University Malaysia, Malaysia | SCOPUS H-Index: 8. Google Scholar H-Index: 11)
- 6. Dr. Wahyu Caesarendra (Opole university, **Poland**) | SCOPUS, H-Index:24 | Google Scholar H-Index:54)
- 7. Muhammad Lahandi Baskoro (Conventry University, United Kingdom) | SCOPUS, H-Index :1 | Google Scholar H-Index :4)

Section Editor

- 1. Assoc.Prof.Silvester Dian Handy Permana,Ph.D (Universitas Trilogi, Indonesia | SCOPUS H-Index: 3, Google Scholar H-Index :8)
- 2. Ade Syahputra, S.T., M.Inf.Comm.Tech.Mgmt (Universitas Trilogi, Indonesia | SCOPUS H-Index :1, Google Scholar H-Index :5)
- 3. Erneza Dewi Krishnasari., S.Ds., M.Ds (Universitas Trilogi, Indonesia | SCOPUS H-Index: 1, Google Scholar H-Index: 3
- 4. Gatot Tri Pranoto., S.Kom., M.Kom (Universitas Trilogi, Indonesia | Scopus H-Index: 1, Google Scholar)

JOURNAL IDENTITY

Journal Name: JISA (Jurnal Informatika dan Sains)

e-ISSN: 2614-8404, p-ISSN: 2776-3234

Publisher: Program Studi Teknik Informatika Universitas Trilogi

Publication Schedule: June and December

Language: English

APC: The Journal Charges Fees for Publishing

Indexing: EBSCO, DOAJ, Google Scholar, Arsip Relawan Jurnal Indonesia, Directory of Research Journals Indexing, Index

Copernicus International, PKP Index, Science and Technology Index (SINTA, S4), Garuda Index

OAI address: http://trilogi.ac.id/journal/ks/index.php/JISA/oai

Contact: jisa@trilogi.ac.id

Sponsored by: DOI - Digital Object Identifier Crossref, Universitas Trilogi

In Collaboration With: Indonesian Artificial Intelligent Ecosystem(IAIE), Relawan Jurnal Indonesia, Jurnal Teknologi dan Sistem Komputer (JTSiskom)



JISA (Jurnal Informatika dan Sains) is Published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.

Journal Profile

Last update: 10 March 2023 Number of documents: 91 Total citations (all years) = 109 h-Index: 5, i10-Index: 2 Google Scholar URL: Click Here

Submit an Article

Article Processing Charge

Focus and Scope

Author Guidelines

Article Templates

Peer Review Process

Publication Frequency

Open Access Policy

Copyright Agreement

Editorial Team

Peer-Reviewers

Collaboration

Statistics

Indexing

Scopus Citation Analysis









#2103 Review 15/10/25, 21.09



номе

JISA(Jurnal Informatika dan Sains)

Journal of Informatics and Science

SEARCH

p-ISSN: 2776-3234

CATEGORIES

ANNOUNCEMENTS

PUBLICATION ETHICS

ARCHIVES

Home > User > Author > Submissions > #2103 > Review

USER HOME

#2103 Review

ABOUT

Submission

Authors Joko Riyono, Christina Eni Pujiastuti, Supriyadi Supriyadi, Dody Prayitno, Aina Latifa Riyana Putri 🕮

CURRENT

Decision Tree for Determining Hospital Treatment for Covid-19 Patients Based on Hematology Title

Parameters Using the C5.0 Algorithm

Articles Section

Editor Wahyu Caesarendra 🕮

Peer Review

Round 1

Review Version 2103-5357-1-RV.DOCX 2025-03-10

Initiated 2025-03-10 Last modified 2025-03-10

Uploaded file Reviewer A 2103-5358-1-RV.DOCX 2025-03-10

Reviewer B 2103-5359-1-RV.DOCX 2025-03-10

Editor Decision

Accept Submission 2025-03-10

Editor/Author Email Record 3 2025-03-10 Notify Editor

2103-5360-1-ED.DOCX 2025-03-10 **Editor Version**

Author Version None

Upload Author Version Choose File no file selected

Upload

JOURNAL IDENTITY

Journal Name: JISA (Jurnal Informatika dan Sains)

e-ISSN: 2614-8404, p-ISSN: 2776-3234

Publisher: Program Studi Teknik Informatika Universitas Trilogi

Publication Schedule: June and December

Language: English

APC: The Journal Charges Fees for Publishing

Indexing: EBSCO, DOAJ, Google Scholar, Arsip Relawan Jurnal Indonesia, Directory of Research Journals Indexing, Index

Copernicus International, PKP Index, Science and Technology Index (SINTA, S4), Garuda Index

OAI address: http://trilogi.ac.id/journal/ks/index.php/JISA/oai

Contact: jisa@trilogi.ac.id

Sponsored by: DOI - Digital Object Identifier Crossref, Universitas Trilogi

In Collaboration With: Indonesian Artificial Intelligent Ecosystem(IAIE), Relawan Jurnal Indonesia, Jurnal Teknologi dan Sistem Komputer (JTSiskom)



JISA (Jurnal Informatika dan Sains) is Published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.

Journal Profile

INDEXING

Last update: 10 March 2023 Number of documents: 91 Total citations (all years) = 109 h-Index: 5, i10-Index: 2 Google Scholar URL: Click Here

HARDCOPY

Submit an Article

Article Processing Charge

Focus and Scope

Author Guidelines

Article Templates

Peer Review Process

Publication Frequency

Open Access Policy

Copyright Agreement

Editorial Team

Peer-Reviewers

Collaboration **Statistics**

Indexing

Scopus Citation Analysis











Decision Tree for Determining Hospital Treatment for Covid-19 Patients Based on Hematology Parameters Using the C5.0 Algorithm

Joko Riyono¹, Christina Eni Pujiastuti², Supriyadi³, Dody Prayitno⁴, Aina Latifa Riyana Putri⁵

1,2,3,4 Teknik Mesin, Fakultas Teknologi Industri, Universitas Trisakti

⁵Sains Data, Fakultas Informatika, Telkom University

Email: jokoriyono@trisakti.ac.id¹, christina.eni@trisakti.ac.id², supri@trisakti.ac.id³, dodyprayitno@trisakti.ac.id⁴,

ainaqp@telkomuniversity.ac.id⁵

Abstract — The rapid spread of the COVID-19 disease, which occurred globally from late 2019 to the early 2020s, significantly impacted communities worldwide, requires early detection of COVID-19 which is very important for patients and also the people around them to be able to fight the COVID-19 pandemic. Therefore, a classification analysis will be carried out to make decisions regarding determining COVID-19 patients who do not require hospitalization or who require Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in hospitals based on hematology parameters from the Machine Learning Repository. Kaggle Dataset uses the C5.0 algorithm assisted by Rstudio software. It is also known that because the data contains missing data, it is also necessary to handle missing data using the Mean Method assisted by SPSS software. Performance evaluated using the Confusion Matrix method produces an accuracy value of 78% which is considered quite good, where testing with the C5.0 Algorithm uses a training and testing data ratio of 40:60. This research simplifies and speeds up medical decision-making, improving patient management. With COVID-19 declining, the method can be applied to enhance healthcare systems' accuracy and efficiency in handling other diseases or emergencies, ensuring better preparedness for future challenges.

Keywords - Early Detection, Classification, Missing Data, Confusion Matrix

I. INTRODUCTION

Coronavirus is part of a group of viruses that generally infect animals but can adapt and eventually transmit to humans. In humans, this virus typically attacks the respiratory system, ranging from mild symptoms like the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS), which emerged in 2012, and Severe Acute Respiratory Syndrome (SARS) in 2002. The latest type of coronavirus was discovered in humans in Wuhan, China, in December 2019. This virus is named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and causes the disease COVID-19. Due to its rapid global spread, the World Health Organization (WHO) declared this disease a pandemic on March 11, 2020.Clinical symptoms of COVID-19 infection can vary from asymptomatic (no symptoms) to fever, cough, runny nose, fatigue, sore throat, and severe conditions (e.g., acute respiratory distress syndrome [ARDS], acute heart injury, and kidney injury) [1]. As of April 19, 2022, the number of COVID-19 cases globally reached 504,571,336. Preventive measures have also been implemented by governments and WHO to help reduce COVID-19 cases, such as requiring activities to be accompanied by strict health protocols as part of daily life, travel restrictions between countries and cities, requiring vaccination, and RT-PCR (Real Time PCR COVID-19) tests as tools to detect the presence of the COVID-19 virus in the body. Furthermore, healthcare systems have been established to detect, test, isolate, treat each case, and track every contact. Preventive measures

play a major role in reducing COVID-19 cases when protocol therapy is applied from the early stages [2]. Therefore, early detection of COVID-19 is crucial for patients and those around them to help prevent a resurgence of the pandemic. When patients receive timely and appropriate care, those around them are also protected. COVID-19 is a systemic infection that significantly impacts the hematopoietic and hemostasis systems, leading to several cardiovascular complications [1], [3], [4]. Research indicates [1] that hematological indices are associated with disease severity and can contribute to decision-making for predicting whether a COVID-19 patient will require ICU admission upon hospital entry. Several hematological abnormalities have also been identified, with significant changes in hematological parameters observed in patients with severe COVID-19 requiring hospitalization and ICU care [5]. The hematological consequences of COVID-19 infection must be utilized by researchers and medical personnel to initiate new treatment approaches or breakthroughs in managing COVID-19 infection.

e-ISSN: 2614-8404

p-ISSN:2776-3234

Therefore, a classification analysis will be conducted to make decisions regarding which COVID-19 patients do not require hospitalization or need care in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital, using a Kaggle Dataset and based on hematological parameters and the C5.0 algorithm. Since the dataset contains missing data, a method to handle missing data using the Mean Method will also be applied. The goal is to facilitate and accelerate the work of medical



personnel, ensuring that COVID-19 patients receive timely and appropriate treatment, ultimately reducing COVID-19 cases in a population.

A. C5.0 Algorithm

The C5.0 algorithm is a data mining method that operates using a decision tree structure. This algorithm is a further development of the ID3 and C4.5 algorithms, with improved efficiency and better capability in categorizing data into appropriate groups. C5.0 is known for its superior performance in solving classification problems [6]. This algorithm has already been applied in various ways [7], such as to analyze the perceived stress levels of healthcare workers treating COVID-19 patients during the early stages of the pandemic in northeastern Mexico. The aim was to understand and categorize their stress levels, producing a visualization model to help identify stress risks and potential mental health issues; [8] as a decision-making tool and comparing performance in breast cancer diagnosis using C5.0 Algorithm and Boosting method.

The Mean method is the most common imputation technique, where missing data in a variable is replaced with the average value of all available data for that variable [9]. A decision tree is a structure resembling a flow diagram shaped like a tree, where each internal node represents a test on an attribute, each branch indicates the test result, and each leaf node represents a class or class distribution that is the final outcome of the test. The C5.0 algorithm is an enhancement of earlier decision tree algorithms developed by Ross Quinlan in 1987, specifically ID3 and C4.5. ID3 evolved into C4.5, which can handle both discrete and continuous attributes. C4.5 was further developed into C5.0 to address certain weaknesses, such as overlapping when handling large amounts of data, which increases decisionmaking time. C5.0 offers higher accuracy, faster decisionmaking, and more efficient memory usage compared to its predecessors.

The tree-building process in the C5.0 algorithm is similar to that of the C4.5 algorithm. However, while the C4.5 algorithm stops after calculating information gain, the C5.0 algorithm continues by calculating the gain ratio using the obtained information gain and entropy. Therefore, the calculations in the C5.0 algorithm involve several attributes, including entropy, information gain, and gain ratio. The C5.0 algorithm can select attributes based on the highest gain ratio.

The equation for calculating entropy is:

$$Entropy(S) = \sum_{j=1}^{k} -p_j \log_2(p_j)$$
 (1)

Where:

S= Set of cases

k = Number of partitions of S

 $p_i = Propotions of S_i to S$

Next, to obtain the Information Gain calculation, the following equation is used:

Information Gain(S, A) = Entropy(S)-
$$\sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i)$$
 (2)

Where:

S = Set of cases

A = Attribute

m = Number of categories in variable A

|Si| = Number of cases in partition i

|S| =Number of cases in S

Finally, to determine an attribute as a node in the C5.0 algorithm, the Gain Ratio is calculated using the formula:

Gain ratio =
$$\frac{Information Gain (S,A)}{\sum_{i=1}^{m} Entropy(S_i)}$$
 (3)

e-ISSN: 2614-8404

p-ISSN:2776-3234

Where:

Gain(S,A) = Gain value of a variable

Si = Entropy value in a variable

The Gain Ratio calculation simplifies the decision tree produced by C5.0 compared to the C4.5 algorithm. The tree is built continuously until no further sample subsets can be split.

B. Mean Method for Handling Missing Data

Missing data is a frequent issue in most research studies, typically arising from non-sampling errors. These errors can include:

Interviewer recording errors, where questions may be skipped during data collection.

Respondent inability errors, where participants fail to provide accurate responses due to misunderstanding the question, experiencing fatigue, or losing interest.

Respondent unwillingness errors, where individuals choose not to answer sensitive questions related to topics such as income, age, weight, or legal history, leading to incomplete responses or abandonment of the survey.

One of the most straightforward and commonly used approaches to address missing data is the Mean imputation method. This technique involves replacing missing values in a dataset with the average of the available values. However, this method is only suitable for numerical data.

The formula used to calculate the mean for imputing missing data is as follows:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Where:

 $\bar{\mathbf{x}} = \mathbf{Mean}$ (average value)

n = Total number of data points

 $x_i = Individual data points$

By applying this formula, missing data points are replaced with the calculated mean of the existing values.

C. The Confusion Matrix

The Confusion Matrix is a useful tool for assessing the accuracy of a classification model by comparing predicted values with actual outcomes. It is applicable to both binary and multi-class classification problems and consists of four key values:

True Positive (TP): The number of cases that are correctly predicted as positive.

True Negative (TN): The number of cases that are correctly predicted as negative.

False Positive (FP): The number of cases incorrectly predicted as positive when they are actually negative.

False Negative (FN): The number of cases incorrectly predicted as negative when they are actually positive.

These values allow for the calculation of Accuracy, which measures how well the model's predictions match the actual values.

The formula for Accuracy is:
$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$
(5)



II. RESEARCH METHODOLOGY

In this study, the data to be used is secondary data from the Kaggle Dataset, "Diagnosis of COVID-19 and its clinical spectrum",

https://www.kaggle.com/datasets/einsteindata4u/covid19

[10], created by Hospital Israelita Albert Einstein in São Paulo, Brazil. This dataset contains anonymous data from 5,644 patients at Hospital Israelita Albert Einstein in São Paulo, Brazil, including information such as patient ID, patient age, and other details. Additionally, samples collected from each patient who underwent RT-PCR testing for SARS-CoV-2 (both positive and negative for COVID-19), supplementary laboratory tests during hospital visits (such as hematocrit, hemoglobin, urineurobilinogen, etc.), as well as the patients' COVID-19 care outcomes, are included. The latter includes whether the patient required care in the Regular Ward, Semi-Intensive Unit, Intensive Care Unit (ICU), or did not require hospitalization at all.

The Regular Ward is a hospital room with more than two beds, intended for patients who no longer require close monitoring and have a low level of dependency, where patients typically begin mobilizing in preparation for discharge.

The Semi-Intensive Unit is designated for patients who still require close monitoring but at a lower level than in the ICU.

The Intensive Care Unit (ICU) is designed for patients who require intensive monitoring and medical support.

In this study, classification analysis will be conducted to determine whether COVID-19 patients do not require hospitalization or if they need care in the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) based on hematological parameters using the C5.0 algorithm with a focus on accuracy. Since the dataset contains missing data, the Mean Method will be applied to handle the missing values. Therefore, several relevant variables that have a correlation with the study's objectives will be selected from the entire dataset.

The variables used in this study are as follows:

Table 1. Variable Identification

No.	Variable	Type
1	Hematocrit	Numerical
2	Hemoglobin	Numerical
3	Platelets	Numerical
4	Mean Platelet Volume (MPV)	Numerical
5	Red Blood Cells	Numerical
6	Lymphocytes	Numerical
7	Mean Corpuscular Hemoglobin Concentration	Numerical
	(MCHC)	
8	Leukocytes	Numerical
9	Basophils	Numerical
10	Mean Corpuscular Hemoglobin (MCH)	Numerical
11	Eosinophils	Numerical
12	Mean Corpuscular Volume (MCV)	Numerical
13	Monocytes	Numerical
14	Red Blood Cell Distribution Width (RDW)	Numerical

In Table 1, the selection of variables is based on several previous studies. For example, in [11], hematocrit can significantly predict the risk of ICU admission in COVID-

19 patients in Iran using multivariable analysis. Study [12] discusses red blood cells less frequently in the pathogenesis of COVID-19, but some studies have considered hemoglobin levels, which are also a major constituent of red blood cells. [13] shows that a decrease in hemoglobin levels in COVID-19 patients is associated with the severity of the disease. There is also a relationship between platelet levels in hospitalized patients and high severity of COVID-19 [14]. The MPV value was found to increase by 6.3% in COVID-19 patients with high severity [15]. Study [16] supports the hypothesis that lymphopenia (a condition when lymphocyte levels are low) can be a prognostic factor in determining the clinical course and severity of the disease in patients hospitalized due to COVID-19. High MCV or low MCHC was found in COVID-19 patients with high or critical severity [17]. Leukopenia (a condition when leukocyte levels are low) has also been reported in several studies, ranging from 28.1% to 68.1%, depending on the severity of the disease and underlying conditions, indicating a possible relationship between the severity of leukopenia and the severity of COVID-19 [12]. Severe COVID-19 cases are typically characterized by low lymphocyte counts, high leukocyte counts, increased neutrophil-lymphocyte ratio (NLR), as well as decreased percentages of monocytes, eosinophils, and basophils [18]. Study [19] found that RDW (Red Cell Distribution Width) could serve as an indicator for predicting the prognosis of COVID-19 patients experiencing severe conditions. A total of 96.4% and 90% of all COVID-19 patients showed low MCH and hemoglobin levels [20].

e-ISSN: 2614-8404

p-ISSN:2776-3234

The stages of data analysis in this study are as illustrated in the following flowchart:

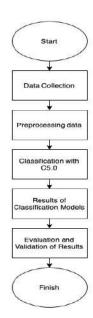


Figure 1. Data Analysis Method Flowchart

As shown in Figure 1, this study begins with the following stages:

Data Collection

This stage describes how and from where the data for this research is obtained. The data will be saved in files with a



.xlsx extension. Data Preprocessing The initial data processing will include Data Selection, which involves selecting data from a larger dataset. The selected data will be used for data mining processes, specifically classification. After that, corrections will be made for errors in the data; in this study, missing values will first be filled using the Mean Method with the assistance of SPSS software. Furthermore, before classification, the data will also be divided into Training and Testing datasets. Classification C5.0This stage involves classifying the data that has been processed in the Data Preprocessing stage using the C5.0 Algorithm **RStudio** with software. Classification Results The classification results will take the form of a decision tree to determine whether COVID-19 patients require hospitalization or need to be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in parameters. the hospital based on hematological Evaluation and Validation of Results In this stage, the classification results will be evaluated

III. RESULTS AND DISCUSSION

using Confusion Matrix measurements

The data mentioned in the Research Method section will undergo Data Selection by choosing several attributes that are relevant to this study. Therefore, from 5.644 patients, 83 COVID-19 positive patients who underwent blood tests upon hospital admission were selected. It is also known which ward these patients were placed in at the hospital. Below is the data used in this study.

PW	91	CU	HC	IME	PL.	HPY	PBC	IFC	MORC	UC	BEP	HOR	BNP	MOV	MME	MA
1	Ū	1	1,99283822	1,7921,8763	-034054761	141938762	165347571	-0.04538327	-14528949	-142018704	13052879	-1.84224524	-0.49889264	-139611351	19833898	19(7)(4))
1	Ū	1	4.6593911	-0.3902/1982	-0.71840131	-0.43889694	-0.56734953	0.98541381	1241457	-88293191	-114114975	1,33498943	-0.66685017	0.22621025	-0.45662277	-1.97889912
1	0	1	43120166	-0.949003	-012790043	-0.11151718	-065621017	-0,8999574	-1.4468135	-196840698	4.52922559	11/218195	017589746	081711755	153312792	134794807
1	0	1	-15180235	-0.27236297	-021592958	0.45944902	-052585816	-0.45777732	19411960	4572547	4.22976651	143953813	-0.71988952	00661462	259738313	-18(19544
1	0	1	14942888	0,729532	-0.74352556	0.195820	0.59565055	4.63688731	134373529	-1.6066390	4.22976651	112590917	-011913823	-0.01417391	0.88281063	-171953947
1	0	I	4.11525339	-01/07/52/12	-110732562	0.91812148	-055081881	163078083	-1553/1511	-1.33573556	4 22576651	1,75316112	0.47001304	110760088	0.72529171	161331779
1	0	1	1,85865	123077541	-0.60534567	0.5736427	0.96583214	-118274617	-0.15416568	0.6811687	114114975	1.2827316	-0.83550740	0.41654644	-0.74558812	-127124712
1	0	1	118761554	1,481,40811	182384102	0.84725538	11423950	0.96657312	189739799	-113537443	122976651	138735134	0.83550792	045172354	-0.37762311	-1.97189912
1	0	1	111127963	190671956	-1,6856612	13449034	159769371	0.3665802	13095000	-145300	122176651	-1.00308017	-0.75122093	-0.71511634	-0.14045413	-1.97109912
				F	igure	2. D	ata c	of CC	VID	-19 I	ositi	ve P	atien	ts		

The description for Figure 2 can be found in Table 2. The next step is to convert the attributes into a format that can be processed by the program.

Table 2. Description of Figure 2

Notation	Description
RW	Regular Ward
	(1=yes, 0=no)
SIU	Semi-Intensive Unit
	(1=yes, 0=no)
ICU	Intensive Care Unit
	(1=yes, 0=no)
HMC	Hematocrit
HMG	Hemoglobin
PTL	Platelets
MPV	Mean Platelet Volume (MPV)
RBC	Red Blood Cells
LPC	Lymphocytes
MCHC	Mean Corpuscular Hemoglobin Concentration (MCHC)
LKC	Leukocytes
BSP	Basophils

MCH	Mean Corpuscular Hemoglobin (MCH)
ESNP	Eosinophils
MCV	Mean Corpuscular Volume (MCV)
MNC	Monocytes
RDW	Red Blood Cell Distribution Width (RDW)

e-ISSN: 2614-8404

p-ISSN:2776-3234

The attributes RW, SIU, and ICU had their values converted from discrete to numeric (label) format in the program, resulting in numeric labels as shown in Table 3.

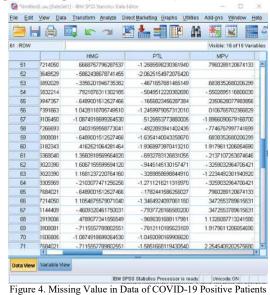
	Table 3. Programming Label Results	
	Description	Numeric
Label	_	
	No hospital care	0
Care	Patient admitted to Regular Ward	1
	Patient admitted to Semi-Intensive Unit	2
	Patient admitted to Intensive Care Unit	3

Figure 3 below shows that after converting the discrete attribute values into labels with numeric values, the data appears simpler and easier to process by machine learning algorithms. The next step is to search for missing values.

Persustan	HME	96	FIL	MPT	800	惠	HIN	LOC	889	HIH	BMF	MCV	HAL	DY
- 1	1993007	LNNSSNI	0.9054768	1400,0007	1634700	0.14030356	0.45289486	440334	19838%	14054585	-(498809)	-0.390113925	1988967	1971403
- 1	4,495(5)00	4.98075811	07080096	4.438893941	-19794513	0.985403888	0.044.4557	4129000	-114043757	039889429	4600006	129/2015	-04661256	4,999902
- 1	432009	1,540000151	0.02760832	4300208	0.660,000	-1.095574	1198035	436465	052525588	0.0228065	LISENSI	13073758	1532753	13079001
1	458000	4.2736971	0.3552990	1.69403	455866	04507531	094,1662	4.575251475	0.2256666	04383305	-0.709(95)(8	19604621	250798390	4,833,9548
1	16943988	1,72950001	0.7452565	1,235062808	1256550	0.50007312	0340735234	1006005	0.22796866	0125903074	-0.1190,900.06	-1114073907	0.88260058	47032346
- 1	4000008	41000019	-1,807123025	130022403	-8.2500LHED	1,51,400,425	0000000	-11373339	-0.22190000	0.75335290	LOOLIES	1100090	072011107	LGDLVZ
1	1385696	1,397340	0.959566	ESTEMBS	1,658025	4.0074072	0154,665	4600004	41408352	039273899	4,000,000	14054544	-074590015	423139789
1	10574.75%	LABOURLE	1,034105	1,347(5)319	1.34335502	0.596575115	18979798	-11357407	02274836	0387361342	4.8888064	-146472537	037703065	4.999902
7	132768	18955591	-1 68805110	178,000,007	1500000000	0.0000000	1.7950900	4.6800001	402209/200	CONTRACTOR	475788	-1700000	40000000	4 920817

Figure 3. Latest Data of COVID-19 Positive Patients

It can be seen that several data entries have missing values, specifically in the MPV variable for patients 52 and 70, as shown in Figure 4. To address this issue, missing value imputation will be performed, considering that missing value imputation is a treatment for outliers in an effort to improve data quality [22. BPS, 2017].



From the process carried out, the results are shown in Figure 5. In the figure below, it can be seen that the MPV variable has two missing values, with a valid count of 81, along with the Mean Standard Deviation Range

variable has two missing values, with a valid count of 81, along with the Mean, Standard Deviation, Range, Minimum Value, Maximum Value, and Sum of the MPV variable.



MON

e-ISSN: 2614-8404 p-ISSN:2776-3234

Statistics

И	Valid	81
	Missing	2
Mear	i	.2752305728
Std. [Deviation	.8972492394
Rang	je	4.599922419
Minin	num	-1.89660907
Maxir	mum	2.703313351
Sum		22.29367640

Figure 5. Output of Step 2: Missing Value Imputation Figure 6 explains that the new variable "MPV_1" will perform imputation on the 2 missing values from 83 data entries using the function "SMEAN(MPV)."

Result Variables

	Result Variable	N of Replaced	Case Number of Value				
		Missing Values	First	Last	N of Valid Cases	Creating Function	
1	MPV_1	2	1	83	83	SMEAN(MPV)	

Figure 6. Output of Step 4: Missing Value Imputation As shown in Figure 7, the MPV values for patients 52 and 70 have been filled or replaced with the value 0.2752306, which is the mean of the MPV variable inserted into the cells containing the missing values.

ile <u>E</u> dit	Yew Da	ta Iransform Analyze	Direct Marketing Graphs 1	Itilities Add-ons	Window He	(p
<u>a</u> .				台 笔		4
:HMC		.9918382167816162		Visible	16 of 16 Varial	oles
		MNC	RDW	MPS	_1	
51	530396	.9616004824638367	- 8019854426383972	,7960	289120674133	•
52	191700	6929816603660583	4.9476857185363770	.2752	305728233891	1
53	547241	1.4868645668029790	-1.0673550367355350	.6838	352680206299	
54	337714	2262306213378906	2594920992851257	- 5502	895116806030	1
55	367834	9616004824638367	2.3824472427368160	.2350	528077983856	1
56	511328	7455080151557922	.0825793221592903	.0106	765702366829	1
57	168530	-1.0081400871276860	.2594920992851257	-1.8966	090679168700	1
68	13987	3252967298030853	-1.2442678213119610	7746	767997741699	1
59	114180	- 1414541900157928	- 6250726580619812	,6838	352680206299	Н
60	121880	3.6404480934143070	4481598734855652	1.9179	511206054690	ı
61	194380	1.2242325544357300	.9671441316604614	-,2137	107253074646	ı
62)11230	.7252317070960999	.1710353046655655	.3259	032964706421	1
63	176990	2727702260017395	- 9789991212844849	-1.2234	492301940920	П
64	799081	3575466573238373	- 9788991212844849	3259	032964706421	1
66	343410	- 0889278352260590	1827902793884277	.7960	289120674133	1
66	139612	.2524939179420471	.5248617529869080	.3472	553789615631	1
67	125543	-1.0081400871276860	.5248617529869080	3472	553789615631	1
68	43347	1.1717061996459960	7135294675827026	1.1326	087713241580	1
69	106690	-1.1131930351257320	2594920992851257	1,9179	511206054690	1
70	194340	2.3535504341125490	3.0016424655914310	.2752	305728233891	
71	169170	-2.0586686134338380	1.2325129508972170	2.2545	409202575680	Ē
Data View	Variable V	New New				

Figure 7. Results of Missing Value Imputation

Before the classification process begins, the data is typically divided into two parts: the training set and the testing set, according to a specific proportion. This separation uses the Train-Test Split method, which serves to evaluate the performance of the machine learning model. The training data is used to build and train the model, while the testing data is used to measure how well the model can predict data it has never seen before. This is important to ensure that the model has good generalization and can perform well on new data. Therefore, the training and testing data will be divided using RStudio software with the

following steps:

• The data that has previously had missing values replaced will be imported in .xlsx file format using the "readxl" library, which has been downloaded beforehand, along with specifying the file name and the folder where the file is stored.

```
## tibble [83 x 15] (53: tbl_df/tbl/data.frame)
## $ Perawatan: num [1:83] 0100000030...
              : num [1:83] 0.992 -0.496 -0.313 -0.519 0.694 ...
: num [1:83] 0.792 -0.396 -0.649 -0.273 0.73 ...
## $ HMC
## $ HMS
## $ PTL
               : num [1:83] -0.3415 -0.7184 -0.0275 -0.2159 -0.7435 ...
               : num [1:83] 1.469 -0.438 -0.102 0.459 0.235 ...
               : num [1:83] 1.653 -0.568 -0.656 -0.515 0.596 .
## $ RBC
## 1 LPC
               : num [1:83] -0.0484 -0.9354 -0.0996 -0.4578 -0.6369 ...
## $ MCHC
               : num [1:83] -0.453 0.244 -1.449 0.941 0.344 ...
               : num [1:83] -0.42 -0.821 -0.968 -0.573 -0.607 ..
## $ BSP
               : num [1:83] 1.304 -1.14 -0.529 -0.224 -0.224 ...
## 5 MCH
               : num [1:83] -1.4422 0.335 0.0214 0.4395 0.1259 ...
               : num [1:83] -0.498 -0.567 0.176 -0.709 -0.119 ...
## $ ESNP
## $ NOV
               : num [1:83] -1.3961 0.2263 0.8071 0.056 -0.0141 ...
## 3 MSC
               : num [1:83] 1.933 -0.457 1.513 2.537 0.883 ...
### $ RDW
               : num [1:83] 0.967 -0.979 0.348 -0.802 -0.714 ...
```

Figure 8. Data Structure

In Figure 8, it can be seen that this dataset consists of 83 entries with 15 variables. Based on the data structure results above, some variable types need to be adjusted according to their characteristics, as described in the previous section. Specifically, the "Care" variable is converted from numeric to factor type with levels "No hospital care," "Patient admitted to Regular Ward," "Patient admitted to Semi-Intensive Unit," and "Patient admitted to Intensive Care Unit" the following using syntax. After changing the "Care" variable to factor data type, Figure 9 shows the structure of the dataset after adjustments have been made to the data structure.

```
## tibble [83 x 15] (S3: tbl_df/tbl/data.frame)
## $ Perawatan: Factor w/ 4 levels "0","1","2","3": 1 2 1 1 1 1 1 1 4 1
## $ HMC
              : num [1:83] 0.992 -0.496 -0.313 -0.519 0.694 ...
## $ HNG
              : num [1:83] 0.792 -0.398 -0.649 -0.273 0.73 ...
              : num [1:83] -0.3415 -0.7184 -0.0275 -0.2159 -0.7435 ...
## $ PTL
## $ NPV
              : num [1:83] 1.469 -0.438 -0.102 0.459 0.235 ...
## $ RBC
              : num [1:83] 1.653 -0.568 -0.656 -0.515 0.596 ...
## $ LPC
              : num [1:83] -0.0484 -0.9354 -0.0996 -0.4578 -0.6369 ...
              : num [1:83] -0.453 0.244 -1.449 0.941 0.344 ...
## $ MCHC
## $ LKC
              : num [1:83] -0.42 -0.821 -0.968 -0.573 -0.607 ...
## $ BSP
              : num [1:83] 1.304 -1.14 -0.529 -0.224 -0.224 ...
## $ NCH
              : num [1:83] -1.4422 0.335 0.0214 0.4395 0.1259 ...
## 1 ESNP
              : num [1:83] -0.498 -0.667 0.176 -0.709 -0.119 ...
## $ HCV
              : num [1:83] -1.3961 0.2263 0.8071 0.066 -0.0141 ...
## $ MNC
              : num [1:83] 1.933 -0.457 1.513 2.537 0.883 ...
              : num [1:83] 0.967 -0.979 0.348 -0.802 -0.714 ...
## $ RDW
```

Figure 9. Data Conversion Structure

To perform the splitting of Training Data and Testing Data from the 83 observation data, this study will allocate 40% as Training Data and 60% as Testing Data randomly. Below is the command to select 33 rows to be stored in the Training Data, with the remaining rows stored in the Testing Data.

```
#Training and Testing Set Preparation (40:60) ###
ic = HC[,-1]
set.seed(1234)
indeks_training_set = sample(83, 33)
input_training_set = ic[indeks_training_set,]
class_training_set = HC[indeks_training_set,]
preparation
input_testing_set = ic[-indeks_training_set,]
```

Figure 10. Data Splitting Syntax



After splitting the data into training and testing sets, the classification process will be carried out using all variables with R software.

The first step is to prepare the packages that will be used during the classification process by installing R packages such as tidyrules, tidyverse, C50, pander, dplyr, and reshape2 using the function "install.packages()." Once the installation process is complete, load the packages into the R session using the function "library()" as shown in the figure below.

```
library(tidyrules)
library(tidyverse)
## -- Attaching packages ------ tidyverse
1.3.1 --
## v ggplot2 3.3.5
## v tibble 3.1.1
## v tidyr 1.1.3
                     v dplyr 1.0.6
v stringr 1.4.0
## v tidyr
## v readr 1.4.0
                     v forcats 0.5.1
## -- Conflicts ----- tidyverse confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()
                   masks stats::lag()
library(C50)
library(dplyr)
## Attaching package: *reshape2
## The following object is masked from 'package:tidyr':
```

Figure 11. Loading Packages in R

And the packages are ready to use.

In Figure 12, "Number of samples" depicts the amount of data used, specifically the Training Data consisting of 33 entries. Meanwhile, "Number of predictors" refers to the number of attributes used, which includes 14 variables with "Care" as the Class in this classification. Using the C5.0 Algorithm, a decision tree with 6 branches will be generated.

```
##
## Call:
## C5.0.default(x = input_training_set, y = class_training_set)
##
## Classification Tree
## Number of samples: 33
## Number of predictors: 14
##
## Tree size: 6
##
## Non-standard options: attempt to group attributes
```

Figure 12. Number of Samples



e-ISSN: 2614-8404

p-ISSN:2776-3234

Figure 13. Output of C5.0 Algorithm Classification

The decision tree obtained from the C5.0 Algorithm is used for determining whether COVID-19 patients require hospitalization or if they should be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters, as shown in the "Decision Tree" section below.



Figure 14. Decision Tree of C5.0 Algorithm Classification

In Figure 14, the decision tree of the C5.0 Algorithm can be interpreted as follows: for example, if a patient undergoes a blood test in the hospital and the ESNP result is > 0.4562533, then the patient can be predicted not to require hospitalization. If both ESNP and MHC values are considered, the patient can be predicted to be admitted to the Regular Ward, and so on. Figure 15 shows a plot of the decision tree from the C5.0 Algorithm.

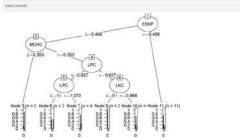


Figure 15. Plot of C5.0 Algorithm Decision Tree



In Figure 13, the "Evaluation on training data" shows an error rate of 18.2% in the classification using the C5.0 Algorithm with the Training Data. Regarding "Attribute usage," it can be observed that there are 4 attributes considered influential in forming the decision tree of the C5.0 Algorithm in this study, with the ESNP attribute being the root or the most important attribute, accounting for 100% usage and so on. Thus, as you move down the tree, the usage of attributes decreases.

After using the Training Data for modeling, the next step is to perform predictions using the C5.0 Algorithm on the Testing Data.

##		hasil_prediksi	0	1	2	3	
##	1	Ø	16	2	2	1	
##	2	1	8	10	2	0	
##	3	2	1	2	0	1	
##	4	3	2	0	0	3	

Figure 16. Output of Syntax 4 for C5.0 Algorithm Predictions

In Figure 16, the confusion matrix from the obtained model is shown. For the 0 category (Patients who do not require hospitalization), there were 0 correct predictions (Patients who do not require hospitalization) based on the previously described hematological parameters, totaling 16 cases. For the 0 category (Patients who do not require hospitalization) predicted as 1 (Patients admitted to the Regular Ward), there were 2 cases, and so on.

To measure the performance of the classification model, the Accuracy will be calculated. Using the formula in equation (5), an Accuracy of 0.78 is obtained. This means that 78% of the patients were correctly predicted as not needing hospitalization or requiring admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU). Therefore, the classification using the C5.0 Algorithm is considered quite good.

IV. CONCLUSION

In this study, a classification analysis was conducted to determine whether COVID-19 patients require hospitalization or admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters using the C5.0 algorithm. A decision tree was obtained for identifying COVID-19 patients who do not require hospitalization or those needing admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) with an accuracy of 78%. The classification using the C5.0 Algorithm is considered quite good.

It is suggested that future research could improve the accuracy by increasing the sample size. It is also hoped that this study can facilitate and expedite the work of medical personnel, enabling COVID-19 patients to receive prompt and appropriate care to help reduce COVID-19 cases in a population.

REFERENCES

e-ISSN: 2614-8404

p-ISSN:2776-3234

- [1] A. Asan *et al.*, "Do initial hematologic indices predict the severity of covid-19 patients?," *Turk J Med Sci*, vol. 51, no. 1, 2021, doi: 10.3906/sag-2007-97.
- [2] M. Khishe, F. Caraffini, and S. Kuhn, "Evolving deep learning convolutional neural networks for early covid-19 detection in chest x-ray images," *Mathematics*, vol. 9, no. 9, 2021, doi: 10.3390/math9091002.
- [3] B. Debuc and D. M. Smadja, "Is COVID-19 a New Hematologic Disease?," Stem Cell Rev Rep, vol. 17, no. 1, 2021, doi: 10.1007/s12015-020-09987-4.
- [4] E. Terpos *et al.*, "Hematological findings and complications of COVID-19," 2020. doi: 10.1002/aih.25829.
- [5] A. Rahman, R. Niloofa, U. Jayarajah, S. De Mel, V. Abeysuriya, and S. L. Seneviratne, "Hematological abnormalities in COVID-19: A narrative review," 2021. doi: 10.4269/ajtmh.20-1536.
- [6] U. S. Aesyi, T. W. Diwangkara, and R. T. Kurniawan, "DIAGNOSA PENYAKIT DISK HERNIA DAN SPONDYLOLISTHESIS MENGGUNAKAN ALGORITMA C5," Telematika, vol. 16, no. 2, 2020, doi: 10.31315/telematika.v16i2.3181.
- [7] E. R. Jorda and A. R. Raqueno, "Predictive model for the academic performance of the engineering students using CHAID and C 5.0 algorithm," *International Journal of Engineering Research and Technology*, vol. 12, no. 6, 2019.
- [8] P. N. Patil, R. Lathi, and V. Chitre, "Comparison of C5 . 0 & CART Classification algorithms using pruning technique," *International Journal of Engineering Research & Technology*, vol. 1, no. 4, 2012.
- [9] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in Classification, Clustering, and Data Mining Applications, 2004. doi: 10.1007/978-3-642-17103-1 60.
- [10] EINSTEIN DATA4U, "Diagnosis of COVID-19 and its clinical spectrum." Accessed: Sep. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/einsteindata4 u/covid19
- [11] A. Sadeghi *et al.*, "COVID-19 and ICU admission associated predictive factors in Iranian patients," *Caspian J Intern Med*, vol. 11, 2020, doi: 10.22088/cjim.11.0.512.
- [12] M. Karimi Shahri, H. R. Niazkar, and F. Rad, "COVID-19 and hematology findings based on the current evidences: A puzzle with many missing pieces," 2021. doi: 10.1111/ijlh.13412.



e-ISSN: 2614-8404 p-ISSN:2776-3234

- Y. Pan et al., "Can routine laboratory tests discriminate SARS-CoV-2-infected pneumonia from other causes of community-acquired pneumonia?," Clin Transl Med, vol. 10, no. 1, 2020, doi: 10.1002/ctm2.23.
- Y.-P. Liu et al., "Combined use of the neutrophil-[14] to-lymphocyte ratio and CRP to predict 7-day disease severity in 84 hospitalized patients with COVID-19 pneumonia: a retrospective cohort study," Ann Transl Med, vol. 8, no. 10, 2020, doi: 10.21037/atm-20-2372.
- [15] G. Lippi, B. M. Henry, and E. J. Favaloro, "Mean Platelet Volume Predicts Severe COVID-19 Illness," 2021. doi: 10.1055/s-0041-1727283.
- J. Wagner, A. DuPont, S. Larson, B. Cash, and A. [16] Farooq, "Absolute lymphocyte count is a prognostic marker in Covid-19: A retrospective cohort review," Int J Lab Hematol, vol. 42, no. 6, 2020, doi: 10.1111/ijlh.13288.
- [17] J. Mao, R. Dai, R. C. Du, Y. Zhu, L. P. Shui, and X. H. Luo, "Hematologic changes predict clinical outcome in recovered patients with COVID-19," Ann Hematol, vol. 100, no. 3, 2021, doi: 10.1007/s00277-021-04426-x.
- [18] C. Qin et al., "Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China," Clin Infect Dis, vol. 71, no. 15, pp. 762-768, Aug. 2020, doi: 10.1093/CID/CIAA248.
- [19] C. Wang et al., "Red cell distribution width (RDW): a prognostic indicator of severe COVID-19," Ann Transl Med, vol. 8, no. 19, 2020, doi: 10.21037/atm-20-6090.
- [20] S. M. Attiyah, H. M. Elsayed, J. A. Al Mughales, A. B. Moharram, and M. A. Fattah, "Critical cases of COVID-19 patients can be predicted by the biomarkers of complete blood count," Indian J Sci Technol, vol. 13, no. 48, pp. 4739-4745, Jan. 2020, doi: 10.17485/IJST/V13I48.2033.



'HFLVLRQ7UHHIRU'HWHUPLQL QJ+RVSLWDO7UHDWPHQWIRU &RYLG3DWLHQWV%DVHGRQ+ HPDWRORJ\3DUDPHWHUV 8VLQJWKH&\$OJRULWKP

by Joko Riyono FTI

Submission date: 07-Nov-2025 12:06PM (UTC+0700)

Submission ID: 2806274622 **File name:** 7._JISA_2.pdf (2.23M)

Word count: 5126 Character count: 27001

Decision Tree for Determining Hospital Treatment for Covid-19 Patients Based on Hematology Parameters Using the C5.0 Algorithm

Joko Riyono¹, Christina Eni Pujiastuti², Supriyadi², Dody Prayitno⁴, Aina Latifa Riyana Putri⁵

1-2-3-6 Teknik Mesin, Fakultas Teknologi Industri, Universitas Trisakti

Sains Data, Fakultas Informatika, Telkom University

Email: jokoriyono@trisakti.ac.id⁴, christina.eni@trisakti.ac.id⁴, supri@trisakti.ac.id⁴, dodyprayitno@trisakti.ac.id⁴,

ainaqp@telkomuniversity.ac.id⁵

Abstract — The rapid spread of the COVID-19 disease, which occurred globally from late 2019 to the early 2020s, significantly impacted communities worldwide, requires early detection of COVID-19 which is very important for patients and also the people around them to be able to fight the COVID-19 pandemic. Therefore, a classification analysis will be carried out to make decisions regarding determining COVID-19 patients who do not require hospitalization or who require Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in hospitals based on hematology parameters from the Machine Learning Repository. Kaggle Dataset uses the C5.0 algorithm assisted by Rstudio software. It is also known that because the data contains missing data, it is also necessary to handle missing data using the Mean Method assisted by SPSS software. Performance evaluated using the Confusion Matrix method produces an accuracy value of 78% which is considered quite good, where testing with the C5.0 Algorithm uses a training and testing data ratio of 40.60. This research simplifies and speeds up medical decision-making, improving patient management. With COVID-19 declining, the method can be applied to enhance healthcare systems' accuracy and efficiency in handling other diseases or emergencies, ensuring better preparedness for future challenges.

Keywords - Early Detection, Classification, Missing Data, Confusion Matrix

I. INTRODUCTION

Coronavirus is part of a group of viruses that generally infect animals but can adapt and eventually transmit to humans. In humans, this virus typically attacks the respiratory system, ranging from mild symptoms like the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS), which emerged in 2012, and Severe Acute Respiratory Syndrome (SARS) in 2002. The latest type of coronavirus was discovered in humans in Wuhan, China, in December 2019. This virus is named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and causes the disease COVID-19. Due to its rapid global spread, the World Health Organization (WHO) declared this disease a pandemic on March 11, 2020.Clinical symptoms of COVID-19 infection can vary from asymptomatic (no symptoms) to fever, cough, runny nose, fatigue, sore throat, and severe conditions (e.g., acute respiratory distress syndrome [ARDS], acute heart injury, and kidney injury) [1]. As of April 19, 2022, the number of COVID-19 cases globally reached 504,571,336. Preventive measures have also been implemented by governments and WHO to help reduce COVID-19 cases, such as requiring activities to be accompanied by strict health protocols as part of daily life, travel restrictions between countries and cities, requiring vaccination, and RT-PCR (Real Time PCR COVID-19) tests as tools to detect the presence of the COVID-19 virus in the body. Furthermore, healthcare systems have been established to detect, test, isolate, treat each case, and track every contact. Preventive measures

play a major role in reducing COVID-19 cases when protocol therapy is applied from the early stages [2]. Therefore, early detection of COVID-19 is crucial for patients and those around them to help prevent a resurgence of the pandemic. When patients receive timely and appropriate care, those around them are also protected. COVID-19 is a systemic infection that significantly impacts the hematopoietic and hemostasis systems, leading to several cardiovascular complications [1], [3], [4]. Research indicates [1] that hematological indices are associated with disease severity and can co 26 bute to decision-making for predicting whether a COVID-19 patient will require ICU admission upon hospital entry. everal hematological abnormalities have also been identified, with significant changes in hematological parameters observed in patients with severe COVID-19 requiring hospitalization and ICU care [5]. The hematological consequences of COVID-19 infection must be utilized by researchers and medical personnel to initiate new treatment approaches or breakthroughs in managing

e-ISSN: 2614-8404

p-ISSN:2776-3234

Therefore, a classification analysis will be conducted to make decisions regarding which COVID-19 patients do not require hospitalization or need care in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital, using a Kaggle Dataset and based on hematological parameters and the C5.0 algorithm. Since the dataset contains missing data, a method to handle missing data using the Mean Method will also be applied. The goal is to facilitate and accelerate the work of medical

COVID-19 infection



personnel, ensuring that COVID-19 patients receive timely and appropriate treatment, ultimately reducing COVID-19 ca. 5 in a population.
A. C5.0 Algorithm

The (24 algorithm is a data mining method that operates using a decision tree structure. This algorithm is a further development of the ID3 and C4.5 algorithms, with improved efficiency and better capability in categorizing data into appropriate groups. C5.0 is known for its superior performance in solving classification problems [6]. This algorithm has already been applied in various ways [7], such as 22 halyze the perceived stress levels of healthcare workers treating COVID-19 patients during the early stages of the pandemic in northeastern Mexico. The aim was to understand and categorize their stress levels, producing a visualization model to help identify stress risks and potential mental health issues; [8] as a decision-making tool and comparing performance in breast cancer diagnosis using C5.0 Algorithm and Boosting method.

The Mean method is the most common imputation technique, where missing data in a variable is replaced with the average value of all available data for that variable [9]. A decision tree is a structure resembling a flow diagram shaped like a tree, where each internal node represents a test on an attribute, each branch indicates the test result, and each leaf node represents a class 3 class distribution that is the final outcome of the test. The C5.0 algorithm is an 5 hancement of earlier decision tree algorithms developed by Ross Quin 21 in 1987, specifically ID3 and C4.5. ID3 evolved into C4.5, which can handle both discrete and hancement of earlier decision tree algorithms developed continuous attributes. C4.5 was further developed into C5.0 to address certain weaknesses, such as overlapping when handling large amounts of data, which increases decisionmaking time. C5.0 offers higher accuracy, faster decisionmaking, and more efficient memory usage compared to its

The tree-building process in the C5.0 algorithm is similar to that of the C4.5 algorithm. However, while 3 C4.5 algorithm stops after calculating information gain, the C5.0 algorithm continues by calculating the gain ratio using the obtained information gain and entropy. Therefore, the calculations in the C5.0 algorithm involve several attributes, including entropy, information gain, and gain ratio. The C5.0 algorithm can select attributes based on the highest gain ratio.

The equation for calculating entropy is:

$$Entropy(S) = \sum_{j=1}^{k} -p_j \log_2(p_j)$$
 (1)

Who7c

S= Set of cases

k = Number of partitions of S

 $p_j = Propotions of S_j$ to S

Next, to obtain the Information Gain calculation, the following equation is used:

Information Gain(S, A) = Entropy(S)-
$$\sum_{i=1}^{m} \frac{|S_i|}{|s|} \times Entropy(S_i)$$
 (2)

Who 7: S = Set of cases

m = Number of categories in variable A

|Si| = Number of cases in partition i

|S| = Number of cases in S

Finally, to determine an attribute as a node in the C5.0 algorithm, the Gain Ratio is calculated using the formula:

$$Gain\ ratio = \frac{Information\ Gain\ (S,A)}{\sum_{i=1}^{m} Entropy(S_i)}$$
(3)

e-ISSN: 2614-8404

p-ISSN:2776-3234

Gain(S,A) = Gain value of a variable

Si = Entropy value in a variable The Gain Ratio calculation simplifies the decision tree produced by C5.0 compared to the C4.5 algorithm. The tree is built continuously until no further sample subsets can be split.

B. Mean Method for Handling Missing Data

Missing data is a frequent issue in most research studies, typically arising from non-sampling errors. These errors can include:

Interviewer recording errors, where questions may be skipped during data collection. Respondent inability errors, where participants fail to

provide accurate responses due to misunderstanding the question, experiencing fatigue, or losing interest.

Respondent unwillingness errors, where individuals choose not to answer sensitive questions related to topics such as income, age, weight, or legal history, leading to incomplete responses or abandonment of the survey.

One of the most straightforward and commonly used approaches to address missing data is the Mean imputation method. This technique involves replacing missing values in a dataset with the average of the available values. However, this method is only suitable for numerical data. The formula used to calculate the mean for imputing missing data is as follows:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Where:

 $\bar{x} = Mean (average value)$

n = Total number of data points xi = Individual data points

By applying this formula, missing data points are replaced with the 11 lculated mean of the existing values.

C. The Confusion Matrix
The Confusion Matrix is a useful tool for assessing the accuracy of a classification model by comparing predicted values with actual outcomes. It is applicable to both bi2 ry and multi-class classification problems and consists of four

True Positive (TP): The number of cases that are correctly predicted as positive

True Negative (TN): The number of cases that are correctly predicted as negative.
False Positive (FP): The number of cases incorrectly

predicted as positive when they are actually negative. False Negative (FN): The number of cases incorrectly

predicted as negative when they are actually positive.

These values allow for the calculation of Accuracy, which measures how well the model's predictions match the actual

values. The formula for Accuracy is:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$
(5)



II. RESEARCH METHODOLOGY

In this study, the data to be used is secondary data from the Kaggle Dataset, "Diagnosis of COVID-19 and its clinical home to be used in the Kaggle Dataset, "Diagnosis of COVID-19 and its clinical home to be used in the spectrum", https: 15 ww.kaggle.com/datasets/einsteindata-du/covid19 [10], created by Hofe tal Israelita Albert Einstein in São Paulo, Brazil. This dataset contains anonymous data from 5,644 patients at Hospital Israelita Albert Einstein in São Paulo, Brazil, including information such as patient ID, patient age, and other details. Additionally, samples collecte 13 from each patient who underwent RT-PCR testing for SARS-CoV-2 (both positive and negative for COVID-19), supplementary laboratory tests during hospital visits (such as hematocrit, hemoglobin, urine-urobilinogen, etc.), as well as the patients' COVID-19 care outcomes, are included. The latter includes whether the patient required care in the Regular Ward, Semi-Intensive Unit, Intensive Care Unit (ICU), or did not require hospitalization at all.

The Regular Ward is a hospital room with more than two beds, intended for patients who no longer require close monitoring and have a low level of dependency, where patients typically begin mobilizing in preparation for discharge. The Semi-Intensive Unit is designated for patients who still

The Semi-Intensive Unit is designated for patients who still require close monitoring but at a lower level than in the U.

The Intensive Care Unit (ICU) is designed for patients who

The Intensive Care Unit (ICU) is designed for patients who require intensive monitoring and medical support.

In this study, classification analysis will be conducted to determine whether COVID-19 patients do not require hospitalization or if they need care in the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) based on hematological parameters using the C5.0 algorithm with a focus on accuracy. Since the dataset contains missing data, the Mean Method will be applied to handle the missing values. Therefore, several relevant variables that have a correlation with the study's objectives will be selected from the entire dataset.

The variables used in this study are as follows:

No.	Variable	Type
1	Hematocrit	Numerical
2	Hemoglobin	Numerical
1	Platelets	Numerical
4	Mean Platelet Volume (MPV)	Numerical
5	Red Blood Cells	Numerical
6	Lymphocytes	Numerical
7	Mean Corpuscular Hemoglobin Concentration (MCHC)	Numerical
8	Leukocytes	Numerical
9	Basophils	Numerical
10	Mean Corpuscular Hemoglobin (MCH)	Numerical
11	Eosinophils	Numerical
12	Mean Corpuscular Volume (MCV)	Numerical
13	Monocytes	Numerical
14	Red Bloed Cell Distribution Width (RDW)	Numerical

In Table 1, the selection of variables is based on several previous studies. For exam [13], in [11], hematocrit can significantly predict the risk of ICU admission in COVID-

19 patients in Iran using multivariable a4 lysis. Study [12] discusses red blood cells less frequently in the pathogenesis of COVID-19, but some studies have considered hemoglobin levels, which are also a major co14 ituent of red blood cells. [13] shows that a decrease in hemoglobin levels in COVID-19 patients is associated with the severity of the disease. There is also a relationship between platelet levels in hospitalized patients and high severity of COVI 4 19 [14]. The MPV value was found to increase by 6.3% in COVID-19 patients with high severity [15]. Study [16] supports the hypothesis that lymphope 18 (a condition when lymphocyte levels are low) can be a prognostic factor in determining the clinical course and severity of the disease in patients hospitalized due to COVID-19. High MCV or low MCHC was found in COVID-19 patients with high or critical severity [17]. Leukopenia (a condition when leukocyte levels are low) has als ⁴ been reported in several studies, ranging from 28.1% to 68.1%, depending on the severity of the disease and underlying conditions, indicating a possible relationship between the severity of leukopenia and the severity of COVID-19 [12]. Severe COVID-19 cases are typically characterized by low lymphocyte counts, high leukocyte counts, increased neutrophil-lymphocyte ratio (NLR), as well as decreased percentages of monocytes, eosinophils, and basophils [18]. Study [19] 23 d that RDW (Red Cell Distribution Width) could serve as an indicator for predicting the prognosis of COVID-19 patients experiencing severe conditions. A total of 96.4% and 90% of all COVID-19 patients showed low

e-ISSN: 2614-8404

p-ISSN:2776-3234

MCH and hemoglobin levels [20]. The stages of data analysis in this study are as illustrated in the following flowchart:



Figure 1. Data Analysis Method Flowchart

As shown in Figure 1, this study begins with the following stages:

Data Collection
This stage describes how and from where the data for this research is obtained. The data will be saved in files with a



Data Preprocessing
The initial data processing will include Data Selection, which involves selecting data from a larger dataset. The selected data will be used for data mining processes, specifically classification. After that, corrections will be made for errors in the data; in this study, missing values will first be filled using the Mean Method with the assistance of SPSS software. Furthermore, before classification, the data will also be divided into Training Testing Classification with This stage involves classifying the data that has been processed in the Data Preprocessing stage using the C5.0 Algorithm with RStudio software. Classification Results The classification results will take the form of a decision tree to determine whether COVID-19 patients require hospitalization or need to be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters. Evaluation and Validation of Results Evaluation In this stage, the classification results will be evaluated using Confusion Matrix measurements

RESULTS AND DISCUSSION

The data mentioned in the Research Method section will undergo Data Selection by choosing several attributes that are relevant to this study. Therefore, from 5.644 patients, 83 COVID-19 positive patients who underwent blood tests upon hospital admission were selected. It is also known which ward these patients were placed in at the hospital. Below is the data used in this study.

9	a	ME	166	n.	501	W.	W.	MX	ut	.81	1601	88	10	MX	W
ī	0	120002	(mps	436507	11980	1895%	414002	94550	140994	10000	14000	0.88361	1,962%	18/09/6	DEM
İ	ţ	115511	CETO	0.004027	040003	-157923	91548	13445	10981	1041475	13490	190300	MENT	16827	117099
1	E	1000	458003	eursus	aptio	498.00	40985	44888	1,810,615	853280	1112875	12800	10072578	issito	11636
1	ć	UNION	022920	0.117000	0692	-100006	397%	1512942	LITSON	12050	18500	9.80051	20000	12,008	10036
1	t	134336	1750	13556	2501	195343	1/58/51	13-0725	10937	1200	(1997	3,00000	0014010.	1163005	130
ŧ	t	42012039	41/20/2	4.005080	198008	-150030	1.00000	-2500301	-, 1007,000	42385	171110	2,000.00	1117109	12012	11175
1	8	1386	120704	18397	1997	1,990,9	13090	DREE	05257	10005	13536	PERSON	0169610	CHESS	92020
1	1	10019	140401	(2000)	0853	192,68	19672	11799	13841	12092	130113	48509	044339	42838	0.0000
1	1	THIRD	168/126	18582	13481	(STREET	13550	LINGH	141336	42000	-OTHE	4500	073315	-113115	artes
	1 1 1 1 1 1	1010	# (1960t) # 6 49501 # 6 49508 # 6 49698 # 6 49698 # 6 49698 # 6 19688 # 6 19688	0 10042-140401 0 10042-140401 0 10066-17552 0 10066-17552 0 10066-17552 0 10066-17552 0 10066-17552	0 189800 07500-03500 0 45000 07500-03500 0 45000 07500-03500 0 45000 07500-03500 0 18000 07500-03500	0 140000 073000 -05537 18000 0 410000 073000 01000 00000 0 410000 000000 010000 00000 0 410000 010000 010000 00000 0 410000 17000 01000 00000 0 410000 17000 01000 01000 0 41000 17000 01000 17000 0 41000 17000 01000 17000 0 11000 17000 01000 01000	0 189800 070300 49537 189800 189675 0 495800 050800 470800 050800 470800 0 405800 450800 050950 05080 470800 0 405800 450800 050800 050800 05080 0 405800 450800 05080 050800 05080 0 405800 450800 05080 05080 05080 0 405800 450800 05080 050800 05080 0 405800 450800 05080 05080 05080 0 105800 150800 050800 050800 05080	0 FEMALE CONTROL HEAVY BERME DESIGN SERVER 0 FEMALE CONTROL DELIVER FORMS - INTROL HEAVY 0 FEMALE CONTROL DELIVER FORMS - INTROL HEAVY 0 FEMALE CONTROL DELIVER CONTROL HEAVY 0 FEMALE CONTROL PROPER FORMS - INTROL HEAVY 0 FEMALE CONTROL FORMS - INTROL HEAVY 0 FEMALE CONTROL FORMS - INTROL HEAVY 0 FEMALE CONTROL HEAVY FORMS - INTROL HEAVY 0 FEMALE CONTROL FORMS - INTROL HEAVY 0 FEMALE	1 HANDE OTTUDO AUSTA HANDE JASSES ALBUST ANDRE \$1 HANDE AUSTA CHARTO CHARTO SORRO AUTORA HANDE HANDE \$1 HANDE AUSTA CHARTO CHARTO SORRO AUTORA HANDE HANDE \$1 HANDE AUSTA CHARTO AUTORA AUTORA AUTORA SONTO HANDE \$1 HANDE SORRO AUTORA CHARTO SONTO HANDE HANDE HANDE \$1 HANDE SORRO AUTORA CHARTO SONTO HANDE HANDE HANDE \$1 HANDE SORRO HANDE AUTORA CHARTO HANDE HANDE HANDE \$1 HANDE SORRO HANDE AUTORA CHARTO HANDE HANDE HANDE \$1 HANDE SORRO HANDE AUTORA CHARTO HANDE H			FIRSTED CHIPPO - RESENT TIRRING TRESS - REPORT - ADDRESS - 1900 - 1000	FIRSTED CHIPPO - NESTO - TARRON IL NESTO - NESTO - APPORT - APPO		1

The description for Figure 2 can be found in Table 2. The next step is to convert the attributes into a format that can be processed by the program.

Table 2	Description o	Figure 2
	Exciteripation of	

Notation	6 escription
RW	Regular Ward
	(1-yes, 0-no)
SIU	Semi-Intensive Unit
	(1-yes, 0-no)
ICU	Intensive Care Unit
	(1-yes, 0-no)
HMC	Hematocrit
HMG	Hemoglobin
PTL	Platelets
MPV	Mean Platelet Volume (MPV)
RBC	Red Blood Cells
LPC	Lymphocytes
MCHC	Mean Corpuscular He 10 John Concentration (MCHC
LKC	Leukocytes
BSP	Basophils

ular Hemoglobin (MCH) Eosinophils puscular Volume (MCV) ESNP MCV MNC Monceytes
RDW Red Blood Cell Distribution Width (RDW)
The attributes RW, SIU, and ICU had their values

e-ISSN: 2614-8404 p-ISSN:2776-3234

converted from discrete to numeric (label) format in the program, resulting in numeric labels as shown in Table 3.

Label	Description	Numeric
	110 hospital care	0
Care	Patient admitted to Regular Ward	1
	Patient admitted to Semi-Intensive Unit	2
	Patient admitted to Intensive Care Unit	3

Figure 3 below shows that after converting the discrete attribute values into labels with numeric values, the data appears simpler and easier to process by machine learning algorithms. The next step is to search for missing values.

Scotte	- 96	186	75.	- 60	185	. 60	100E	100	. 10	101	NP I	101	180	89
1	1,52323	CODES	1.8576	140000	1885%	106825	(8356)	1,0079	1,05234	140304	1.0034	135380	Loadst	disks
112	0.400CK	CHEST	419000	155004	6,20,9600	100400	DATES	460ME	LINCOLL	DADAD	automac.	CONDICT	4.66206	SPREAM
- 6	1 total	-100005	125042	-centre	Challe	00000	(Heath)	1000IT	1000004	1500	Literet	carpose	183290	430760
- 8			12880											
			s.kazet											
			120000											
			LOsse											
			12908											
- 4	133700	100700	1,6600	134660	1100000	19060	1100000	-cataor:	1007059	10000	E2289	4.70300	441944	4YOR I

Figure 3. Latest Data of COVID-19 Positive Patients It can be seen that several data entries have missing values, specifically in the MPV variable for patients 52 and 70, as shown in Figure 4. To address this issue, missing value imputation will be performed, considering that missing value imputation is a treatment for outliers in an effort to improve data quality [22. BPS, 2017].



Figure 4. Missing Value in Data of COVID-19 Positive Patients
From the process carried out, the results are shown in

Figure 5. In the figure below, it can be seen that the MPV variable has t 25 missing values, with a valid count of 81, along with the Mean, Standard Deviation, Range, Minimum Value, Maximum Value, and Sum of the MPV variable



e-ISSN: 2614-8404 p-ISSN:2776-3234

Ctatietice

MPV

N	Valid	81
	Missing	2
Mean	10	.2752305728
Std. I	Deviation	.8972492394
Rang	je	4.599922419
Minin	num	-1.89660907
Maxin	num	2.703313351
Sum		22.29367640

Figure 5. Output of Step 2: Missing Value Imputation
Figure 6 explains that the new variable "MPV 1" will
perform imputation on the 2 missing values from 83 data
entries using the function "SMEAN(MPV)."

Orest Market

	Result Variable	No! Replaced	Case Number of Value			90300
		Missing Values	First	Last	N of Valid Cases	Creating Function
1	VP/_1	- 1	1	83	83	SVEAUNE

Figure 6. Output of Step 4: Missing Value Imputation
As shown in Figure 7, the MPV values for patients 52 and
70 have been filled or replaced with the value 0.2752306,
which is the mean of the MPV variable inserted into the
cells containing the missing values.



Figure 7. Results of Missing Value Imputation 17

Before the classification process begins, the data is typically divided into two parts: the training set and the testing set, according to a specific proportion. This separation uses the Train-Test Split method, which serves 3 evaluate the performance of the machine learning model. The training data is used to build and train the model, while the testing data is used to measure how well the model can predict data it has never seen before. This is important to ensure that the model has good generalization and can perform well on new data. Therefore, the training and testing data will be divided using R Studio software with the

following

• The data that has previously had missing values replaced will be imported in xlsx file format using the "readx!" library, which has been downloaded beforehand, along with specifying the file name and the folder where the file is stored.

Figure 8. Data Structure

In Figure 8, it can be seen that this dataset consists of 83 entries with 15 variables. Based on the data structure results above, some variable types need to be adjusted according to their characteristics, as described in the previous section. Specifically, the "Care" variable is converted from numeric to factor type with levels "No hospital care," "Patient admitted to Regular Ward," "Patient admitted to Semilntensive Unit," and "Patient admitted to Intensive Care Unit" using the following syntax. After changing the "Care" variable to factor data type, Figure 9 shows the structure of the dataset after adjustments have been made to the data structure.

Figure 9. Data Conversion Structure

To perform the splitting of Training Data and Testing Data from the 83 observation data, this study will allocate 40% as Training Data and 60% as Testing Data randomly. Below is the command to select 33 rows to be stored in the Training Data, with the remaining rows stored in the Testing Data.

```
#Training and Testing Set Preparation (40:60)

ic = k([,+1]

set.sec6(1234)

indeks_training_set = sample(83, 33)

input_training_set = %C[indeks_training_set,]

class_training_set = %C[indeks_training_set,]

input_testing_set = ic[-indeks_training_set,]
```

Figure 10. Data Splitting Syntax



After splitting the data into training and testing sets, the classification process will be carried out using all variables with \ensuremath{R} software.

The first step is to prepare the packages that will be used during the classification process by installing R packages such as tidyrules, tidyrerse, C50, pander, dplyr, and reshape2 using the function "install.packages()." Once the installation process is complete, load the packages into the R session using the function "library()" as shown in the figure below.

```
## The following object is masked from 'packagestidyr's ## aniths
```

Figure 11. Loading Packages in R
And the packages are ready to use.
In Figure 12, "Number of samples" depicts the amount of data used, specifically the Training Data consisting of 33 entries. Meanwhile, "Number of predictors" refers to the number of attributes used, which includes 13 variables with "Care" as the Class in this classification. Using the C5.0 Algorithm, a decision tree with 6 branches will be generated.

```
## (all:
## C5.0.default(x = input_training_set, y = class_training_set)
## Classification Tree
## Number of samples: 33
## Number of predictors: 14
## Tree size: 6
## Mon-standard options: attempt to group attributes
```

Figure 12. Number of Samples

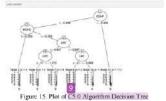
e-ISSN: 2614-8404 p-ISSN:2776-3234

```
E4 [School FOTOR Satisfied] Tub Set 31 (3-26.07-829)
The form () I married from outdined date of the control of the con
```

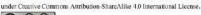
The decision tree obtained from the C5.0 Algorithm is used for determining whether COVID-19 patients require hospitalization or if they should be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters, as shown in the "Decision Tree" section below.



In Figure 14, the decision tree of the C5.0 Algorithm can be interpreted as follows: for example, if a patient undergoes a blood test in the hospital and the ESNP result is > 0.4562533, then the patient can be predicted not to require hospitalization. If both ESNP and MHC values are considered, the patient can be predicted to be admit 51 to the Regular Ward, and so on. Figure 15 shows a plot of the decision tree from the C5.0 Algorithm.



JISA (Jumal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi





In Figure 13, the "Evaluation on training data" shows an error rate of 18.2% in the classification using the C5.0 Algorithm with the Training Data. Regarding "Attribute usage," it can be observed that 12 re are 4 attributes considered influential in forming the decision tree of the C5.0 Algorithm in this study, with the ESNP attribute being the root or the most important attribute, accounting for 100% usage and so on. Thus, as you move down the tree, the usage of attributes decreases.

After using the Training Data for modeling, the next step is [2] to perform predictions using the C5.0 Algorithm on the Testing Data.

##		hasil_prediksi	0	1	2	3	
##	1	0	16	2	2	1	
##	2	1	8	10	2	Ø	
##	3	2	1	2	ø	1	
##	4	3	2	Ø	0	3	

Figure 16. Output of Syntax 4 for C5.0 Algorithm Predictions

In Figure 16, the confusion matrix from the obtained model is shown. For the 0 category (Patients who do not require hospitalization), there were 0 correct predictions (Patients who do not require hospitalization) based on the previously described hematological parameters, totaling 16 cases. For the 0 category (Patients who do not require hospitalization) predicted as 1 (Patients admitted to the Regular Ward),

19 re were 2 cases, and so on.

To measure the performance of the classification model, the Accuracy will be calculated. Using the formula in equation (5), an Accuracy of 0.78 is obtained. This means that 78% of the patients v3 correctly predicted as not needing hospitalization or requiring adf 3 sion to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU). Therefore, the classification using the C5.0 Algorithm is considered quite good.

IV. CONCLUSION

In this study, a classification analysis was conducted to determine whether COVID-19 patients require hospitalization or admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters using the C5.0 algorithm. A decision tree was obtained for identifying COVID-19 patients who do not require hospitalization or those needing admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) with an accuracy of 78%. The classification using the C5.0 Algorithm is considered quite good.

C5.0 Algorithm is considered quite good. It is suggested that future research could improve the accuracy by increasing the sample size. It is also hoped that this study can facilitate and expedite the work of medical personnel, enabling COVID-19 patients to receive prompt and appropriate care to help reduce COVID-19 cases in a

REFERENCES

e-ISSN: 2614-8404

p-ISSN:2776-3234

- A. Asan et al., "Do initial hematologic indices predict the severity of covid-19 patients?," Tark J Med Sci, vol. 51, no. 1, 2021, doi: 10.3906/sag-2007-97
- [2] M. Khishe, F. Caraffini, and S. Kuhn, "Evolving deep learning convolutional neural networks for early covid-19 detection in chest x-ray images," *Mathematics*, vol. 9, no. 9, 2021, doi: 10.3390/math9091002.
- [3] B. Debuc and D. M. Smadja, "Is COVID-19 a New Hematologic Disease?," Stem Cell Rev Rep, vol. 17, no. 1, 2021, doi: 10.1007/s12015-020-09987-4.
- [4] E. Terpos et al., "Hematological findings and complications of COVID-19," 2020. doi: 10.1002/aih.25829.
- [5] A. Rahman, R. Niloofa, U. Jayarajah, S. De Mel, V. Abeysuriya, and S. L. Seneviratne, "Hematological abnormalities in COVID-19: A narrative review," 2021. doi: 10.4269/ajtmh.20-1536.
- 6] U. S. Aesyi, T. W. Diwangkara, and R. T. Kurniawan, "DIAGNOSA PENYAKIT DISK HERNIA DAN SPONDYLOLISTHESIS ALGORITMA C5," *Telematika*, vol. 16, no. 2, 2020, doi: 10.31315/telematika.v16i2.3181.
- [7] E. R. Jorda and A. R. Raqueno, "Predictive model for the academic performance of the engineering students using CHAID and C 5.0 algorithm," International Journal of Engineering Research and Technology, vol. 12, no. 6, 2019.
 - P. N. Patil, R. Lathi, and V. Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique," International Journal of Engineering Research & Technology, vol. 1, no. 4, 2012.
- [9] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in Classification, Clustering, and Data Mining Applications, 2004. doi: 10.1007/978-3-542-17103-1 60.
- [10] EINSTEIN DATA4U, "Diagnosis of COVID-19 and its clinical spectrum." Accessed: Sep. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/einsteindata4 u/covid19
- [11] A. Sadeghi et al., "COVID-19 and ICU admission associated predictive factors in Iranian patients," Caspian J Intern Med, vol. 11, 2020, doi: 10.22088/ciim.11.0.512.
- [12] M. Karimi Shahri, H. R. Niazkar, and F. Rad, "COVID-19 and hematology findings based on the current evidences: A puzzle with many missing pieces," 2021. doi: 10.1111/ijih.13412.



JISA (Jurnal Informatika dan Sains) Vol. 07, No. 02, December 2024

- [13] Y. Pan et al., "Can routine laboratory tests discriminate SARS-CoV-2-infected pneumonia from other causes of community-acquired pneumonia?," Clin Transl Med, vol. 10, no. 1, 2020, doi: 10.1002/ctm2.23.
- [14] Y.-P. Liu et al., "Combined use of the neutrophilto-lymphocyte ratio and CRP to predict 7-day disease severity in 84 hospitalized patients with COVID-19 pneumonia: a retrospective cohort study," Ann Transf Med, vol. 8, no. 10, 2020, doi: 10.21037/atm-20-2372.
- [15] G. Lippi, B. M. Henry, and E. J. Favaloro, "Mean Platelet Volume Predicts Severe COVID-19 Illness," 2021. doi: 10.1055/s-0041-1727283.
- [16] J. Wagner, A. DuPont, S. Larson, B. Cash, and A. Farooq, "Absolute lymphocyte count is a prognostic marker in Covid-19: A retrospective cohort review," Int J Lab Hematol, vol. 42, no. 6, 2020, doi: 10.1111/ijlh.13288.
- [17] J. Mao, R. Dai, R. C. Du, Y. Zhu, L. P. Shui, and X. H. Luo, "Hematologic changes predict clinical outcome in recovered patients with COVID-19," Ann Hematol, vol. 100, no. 3, 2021, doi: 10.1007/s00277-021-04426-x.
- [18] C. Qin et al., "Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China," Clin Infect Dis, vol. 71, no. 15, pp. 763–768, Aug. 2020. doi: 10.1093/ICIJCIAA248
- 762–768, Aug. 2020, doi: 10.1093/CID/CIAA248.
 [19] C. Wang et al., "Red cell distribution width (RDW): a prognostic indicator of severe COVID-19," Ann Transl Med, vol. 8, no. 19, 2020, doi: 10.21037/atm-20-6090.
- [20] S. M. Attiyah, H. M. Elsayed, J. A. Al Mughales, A. B. Moharram, and M. A. Fattah, "Critical cases of COVID-19 patients can be predicted by the biomarkers of complete blood count," *Indian J Sci Technol*, vol. 13, no. 48, pp. 4739–4745, Jan. 2020, doi: 10.17485/UST/V13148.2033.

JISA (Jumal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.



e-ISSN: 2614-8404

p-ISSN:2776-3234

'HFLVLRQ7UHHIRU'HWHUPLQLQJ+RVSLWDO7UHDWPHQWI... &RYLG3DWLHQWV%DVHGRQ+HPDWRORJ\3DUDPHWHUV 8VLQJWKH&\$OJRULWKP

13	% Y INDEX	8% INTERNET SOURCES	10% PUBLICATIONS	7% STUDENT PA	APERS
PRIMARY SO	URCES				
	Submitte tudent Paper	d to University	of Huddersfie	eld	1%
	www.md				1%
N A	Model fo	kholis, Styawat r Soybean Land า", Jurnal Onlin	d Suitability Us	sing C5.0	1%
the state of the s	Rad. "CO' based on many mis	orimi Shahri, Ha VID-19 and her the current ev ssing pieces", Ir ry Hematology	matology findi vidences: A pu nternational Jo	ngs zzle with	1%
\sim	doaj.org				1%
	Submitte tudent Paper	d to University	of Lincoln		1%
/	ournal.u nternet Source	npacti.ac.id			1%
2		010 MONDAY S tensive Care M		October	1%
9	Submitte tudent Paper	d to Universita	s Sebelas Mar	ret	<1%

	Student Paper	<1%
11	Phenikaa University Publication	<1%
12	Pu Song, Xuan Cai, Dan Qin, Qingqing Wang, Xiangwei Liu, Mengmeng Zhong, Linying Li, Yan Yang. "Analyzing psychological resilience in college students: A decision tree model", Heliyon, 2024 Publication	<1%
13	platcovid.com Internet Source	<1%
14	pubmed.ncbi.nlm.nih.gov Internet Source	<1%
15	www.inass.org Internet Source	<1%
16	www.repositorio.ufop.br Internet Source	<1%
17	Submitted to University of Liverpool Student Paper	<1%
18	www.dovepress.com Internet Source	<1%
19	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
20	tudr.thapar.edu:8080 Internet Source	<1%
21	www.futuresmag.com Internet Source	<1%
22	Submitted to California Southern University Student Paper	<1%

23	daten-quadrat.de Internet Source	<1%
24	dokumen.pub Internet Source	<1%
25	dspace.ucuenca.edu.ec Internet Source	<1%
26	www.mlit.go.jp Internet Source	<1%

Exclude quotes

On

Exclude matches

< 10 words

Exclude bibliography On