Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 5, No.1, March 2022, pp. 37 – 43
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Implementation of Decision Tree Algorithm Machine Learning in Detecting Covid-19 Virus Patients Using Public Datasets

[1]Nadiah, [2]Sopian Soim, [3]Sholihin
[1,2,3]Department of Electrical Engineering, Study Program of Telecommunication Engineering,
Politeknik Negeri Sriwijaya
Email: [1]nadiahk25@gmail.com, [2]sopiansoim@gmail.com, [3]sholihin@polsri.ac.id

Article Info

Article history:
Received Apr 10th, 2022
Revised May 20th, 2022
Accepted Jun 6th, 2022

Keyword:
Covid-19 Virus
Classification
Decision Tree Algorithm
Machine Learning
Public Datasets

ABSTRACT

The advancement of Artificial Intelligence (AI) technology has been widely implemented in numerous sectors of daily life. Machine Learning is one of the subfields of Artificial Intelligence. Using statistics, mathematics, and data mining, machine learning is developed so that machines may learn by assessing data without being reprogrammed. At this time the world is on alert for the spread of a popular virus, the corona virus. Coronaviruses are part of a family of viruses caused by diseases ranging from the flu. The disease caused by the coronavirus is known as Covid-19. Therefore, to help identify whether a somebody has coronavirus disease based on certain symptoms, a model is created that can classify people with the covid-19 virus using machine learning. The classification methods utilized in this study are decision trees and large-scale machine learning projects. The study employed Python 3.7 as its programming language and PyCharm as its Integrated Development Environment (IDE). Based on the results, the accuracy rate as expected after conducting various trials is 99%.

Corresponding Author:
Nadiah,
Departement of Electrical Engineering, Study Program of Telecommunication Engineering,
Politeknik Negeri Sriwijaya
Jl. Srijaya Negara, Bukit Besar, Kecamatan Ilir Barat I, Kota Palembang, Sumatera Selatan.
Email: nadiahk25@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v5i1.17054

1. INTRODUCTION

The first time the coronavirus appeared and attacked humans in Wuhan province, China. Symptoms of coronavirus are cough, shortness of breath, fatigue, fever, and no appetite. The coronavirus develops so quickly that it can result in more severe infections and even organ failure. This condition generally occurs in patients who previously had health problems. Covid-19 has spread to 196 countries, positive confirmed cases there are 414,179 and confirmed cases died as many as 18,440 [1].

On March 11, 2020, the World Health Organization (WHO) announced the coronavirus outbreak as a pandemic. This coronavirus outbreak has seriously shaken the world community, almost 200 countries in the world have been infected by this virus. Lockdown and social distancing are methods governments use to restrict the spread of the Covid-19 virus by disrupting the chain of transmission. In this way, every community will not become infectious or infected because it does not come into contact with anyone so that the rate of spread can decrease [2].

This study attempts to assist in determining whether somebody has coronavirus disease based on a set of standardized specified symptoms. These symptoms are referred to recommendations from the WHO and the Ministry of Health and Family Welfare, India. These results or analyses should be considered medical advice. From the results of this study, a decision tree machine learning model was obtained in classifying patients with the Covid-19 virus and the accuracy of the decision tree in making decisions and classifying patients with the Covid-19 virus.

The study [3] described the C4.5 algorithm to diagnose and classify diseases, including analysis of medical records using the C4.5 algorithm to discover disease group trends. The findings of performing the C4.5 algorithm can be used to examine community disease patterns. Another study was conducted on the C4.5 algorithm in diagnosing pneumonia. The findings of the study indicated that the C4.5 algorithm successfully modeled the decision tree with 10 pneumonia rules [4].

Research [5] explains about classifying Covid-19 surveillance datasets using decision trees. This Covid-19 surveillance dataset was obtained through the UCI machine learning repository, a public data repository. This research showed a better accuracy with the accuracy rate obtained employing the decision tree algorithm of 65%.

Research [6] was done to forecast student academic performance during the Covid-19 pandemic so that the findings might affect school policies. The investigated data comprises absenteeism characteristics, assignments, daily repetitions, and test scores that influence the choice criteria for student academic performance in online lessons. The criteria for academic performance decisions consist of "Satisfactory" and "Unsatisfactory".

From the explanation above, there are many issues that should be overcome in limiting the transmission of the Covid-19 virus. Lack of community-wide information and awareness regarding the prevention and control of Covid-19 is a contributing cause to the virus's rapid spread. Then, there are issues in classifying Covid-19 patients, early detection or early diagnosis, infection prevention and control, risk communication, and community empowerment. Many of these issues are due to the absence of application and testing of artificial intelligence in the public domain for diagnosing the Covid-19 virus. This diagnostic can be tested and enhanced using one of the intelligent systems with the decision tree algorithm. Because the Covid-19 virus dataset is used a lot, machine learning is needed in compiling the decision tree. By entering the data then we will ask the machine (machine learning) to make the tree.

From the problems that have been outlined earlier, this study aims to help identify whether an individual has coronavirus disease based on several symptoms. The Covid-19 dataset in this study was taken from the kaggle website. Kaggle is a site and platform for creating the best models for analyzing and predicting datasets. Kaggle data is public data that has been evaluated in all areas. In the health industry, there are numerous statistics, like Covid-19 case data that are currently a research trend in recent years.

2. RESEARCH METHOD

2.1. Literature Review

2.1.1. Covid-19 Virus

In humans, the coronavirus was discovered when in December 2019, an extraordinary incident occurred in Wuhan, China. Coronavirus Disease-2019 (Covid-19) is also known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2) [7]. But cases are on the rise in South Korea, Italy and Iran. The corona outbreak is increasingly widespread globally and poses a serious threat to the global economy [8]. Even WHO stated that the new coronavirus (Covid-19) can persist for many hours or even days at temperatures between 26 and 27 degrees Celsius. When news related to the Corona virus was first heard, many countries were panicked about the spread of the virus, but there were also those who responded casually to the outbreak of this virus [9]. More than 1,100,000 cases have been reported in more than 200 countries and territories as of April 4, 2020, resulting in more than 58,900 deaths. More than 226,000 people have recovered [10].

2.1.2. Data Mining

Data mining is a set of procedures designed to extract the added value of a data set in the form of previously unknown information. Therefore, data mining becomes an increasingly fundamental technique for turning these data into information. Data mining is a data analysis technique based on statistical applications that aim to extract information. With data mining, large amounts of data can be used as other useful information. Data mining can do jobs such as estimating and classifying group data [11].

Data mining is the process of utilizing specific techniques and methods to discover interesting patterns or information within a set of data. Data mining techniques, methods, and algorithms vary greatly. Knowledge Discovery in Database (KDD) as a whole determines to a great extent the method or algorithm that should be utilized [12].

Data mining is the process of identifying new relevant connections, patterns, and trends by sifting through enormous amounts of data recorded in databases using pattern recognition technologies and statistical and mathematical methods [13].

2.1.3. Clasiffication

One of the fundamental functions of data mining is classification. Classification is included in supervised learning since the classification process includes a data-driven learning step. Algorithms employ

this method to identify patterns in data that can be applied to new data whose categories are unknown. Classification techniques are frequently employed in the actual world, as well as in medicine, education, construction engineering, and other professions [14].

The classification process begins with the first data utilized for algorithmic learning or training. Clearly, the training data in question are properties or characteristics of labels. The label denotes the final output of data that will be computed by an algorithm. For instance, there is student registration data with the labels registration/not registered. The program will analyse these data to identify patterns, rules, or new information. Later, this new pattern or information can be utilized to anticipate whether a new record with an unknown label exists. The algorithm's precision varies according on the type of data it analyses [15].

In order to do a classification, historical data is required, which will then be turned into a new rule or body of knowledge. The primary categorisation of problems is as follows [16]:

1. The classification problem departs from the available training data.
2. The training data will be processed using the classification algorithm.
3. The classification problem is solved by the generation of knowledge represented as diagrams, rules, or knowledge.

2.1.4. Machine Learning

Machine learning is a method used to develop programs that can learn from data. Contrary to static computer systems, machine learning programs are designed to teach themselves. How machine learning programs are taught is analogous to how humans are taught, specifically by example. To decide the answers to the following questions, machine learning will evaluate patterns from the analyzed cases. Machine learning is like creating a program that is biased to guess black boxes that have an unknown function formula. The black box is given an input and will produce a specific output. From the input and output data obtained, the program will guess the function formula that is closest to accuracy [17].

Machine learning is machine learning that is very helpful in solving problems, making it easy to do things. In the hospital or health sector, machine learning makes it easy to do something, for example, doctors can diagnose heart disease in a fast time without taking a long time [18]. Machine learning (ML) is employed to train machines how to more efficiently handle data. The objective of machine learning is to gain knowledge from data. Numerous studies have investigated ways to make machines learn without being explicitly programmed [19].

2.1.5. Decision Tree

A tree is a data structure composed of nodes and ribs (edges). There are three types of nodes on a tree: the root node (root/node), the branching/internal node (branch/internal node), and the leaf node (leaf/node). The decision tree is a simplification of classification techniques for a finite number of classes, with internal nodes and root nodes labeled as attribute names, ribs labeled as possible attribute values, and leaf nodes labeled as different classes [20].

Decision tree is one of the machine learning techniques (machine learning) that uses hierarchical sequential structure classification rules by partitioning the training dataset recursively [21]. Decision tree is a flowchart structure shaped like a tree, where each inner node signifies a test of an attribute, with the resulting branch showing the test results, and the leaf node representing the distribution of class [22].

2.2. Methodology

The design of the research begins with the design of the block diagram system as a whole. Flowchart is one of the most important parts in the design of a study, because from the flowchart this series can be known how the research works as a whole. So that the entire block of research diagrams will produce a system that can be functioned. In the flowchart above explains the stages carried out in the research methods carried out.

1. Dataset
   People with the Covid-19 virus to be classified are the severity consisting of mild, moderate, non-existent and severe. The total amount of data is 316800 data. Using public datasets from its kaggle link website https://www.kaggle.com/iamhungundji/covid19-symptoms-checker?select=Cleaned-Data.csv

2. Split Data Training and Testing
   At this stage, the data is first divided into training data and testing data. Data training is a complete set of data containing classes and predictors to be trained so that models can group into the right classes. While data testing is containing new data that will be grouped by the model to find out the accuracy of the model that has been created. Or it can be interpreted that data training is data used to

train a model in studying the patterns or characteristics of a data, while data testing is data used to test models that have previously known the patterns and characteristics of a data.

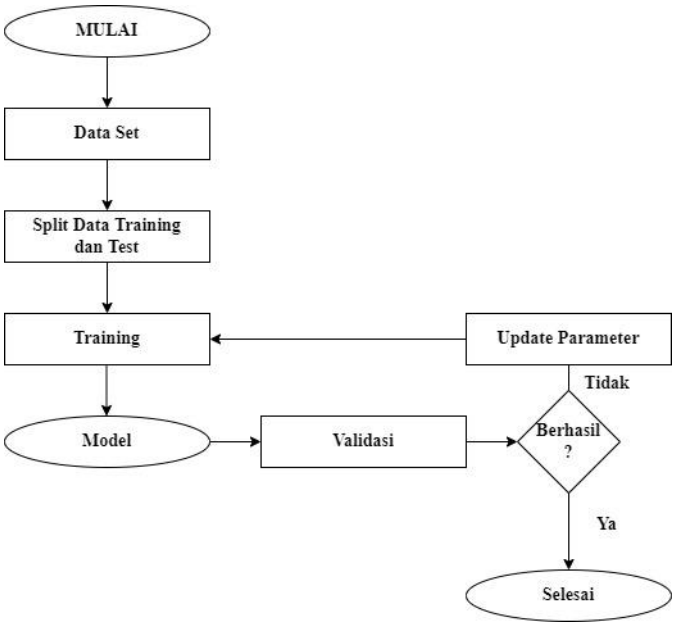

Figure 1. *Flowchart* Program

3. Training
   The training process in machine learning is a machine learning algorithm. The machine learning algorithm will change the parameters on it to adjust to the data provided during the exercise. Just like the human brain, synapses will make changes as humans learn. It is necessary to do training on machine learning to be smart, so machine learning will be trained or learn from the data provided to be able to understand the information on the data.

4. Models
   Use the decision tree function by using the Python programming language. The proposed model/method will be formed from the data that has been processed, and the results of the model processing will be measured with the current model. Decision tree starts from the topmost root, if given a number of test data, for example X where the class of data X is not yet known, then the decision tree will trace from the root to the node and each value of the attribute according to data X is tested whether it is in accordance with the rules of the decision tree, then the decision tree will predict the class of the X list.

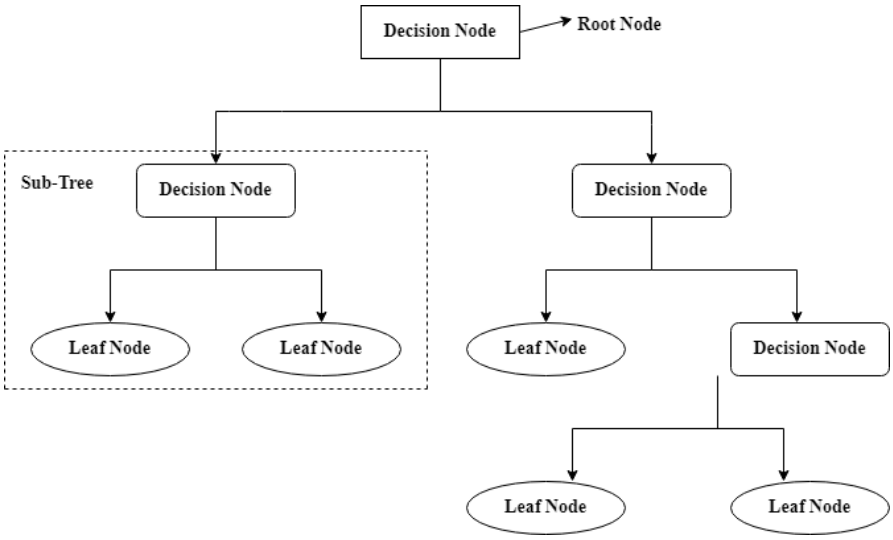

Figure 2. *Decision Tree* Concept

5. Validation
   Validation is used to ensure that the results will be the same when done independently. Validation is the process of activities to assess product design. The validation of the model is aimed at measuring the performance of the model in its ability to make predictions in the data class it tests. So that we can find out the performance of the model and with it can help us to optimize the parameters on the model itself so that the model is much more accurate.

3. RESULTS AND ANALYSIS

In this study, the dataset that will be utilized for training and testing has undergone preprocessing so that it is ready for processing. The classification process for this system uses Python software. Training data and testing data for people with covid-19 using public data from Kaggle websites.

In this data, it will be processed by containing 6 main variables that will have an impact on someone affected by the Covid-19 virus disease or not, while the 6 (six) variables include country, age, symptoms, other symptoms, severity and contact. In this data serves to help identify whether an individual has coronavirus disease based on many earlier symptoms. The dataset contains six variables that influence whether or not a person is affected by coronavirus disease. The description of the variables is provided as follow.

1. Country: List of countries visited by people
2. Age: Age groups classification for each individual, following WHO Age Group Standards
3. Symptoms: Based on WHO, the 5 main symptoms of Covid-19 are fever, fatigue, difficulty breathing, dry cough, and sore throat.
4. Experiencing other symptoms: pain, stuffy nose, runny nose, diarrhea and others.
5. Severity: mild, moderate, severe, none.
6. Contact: Has the individual ever interacted with another Covid-19 patient? (do not know, no and yes)

With all variables of this category, there will be 316800 combinations generated, one for each label in the variable. The results of applying the decision tree algorithm model, obtained some accuracy values as shown in the table below:

Table 1. The accuracy rate of the *decision tree* parameter change

Number of Datasets	Parameter Decision Tree	Parameter Value Decision Tree	Train Accuracy	Test Accuracy
316800	criterion	"entropy"	47%	3.1%
316800	criterion, max depth	"entropy", 10	28%	21%
316800	criterion, max depth	"entropy", 15	38%	11.9%
316800	criterion, max depth	"entropy", 20	46%	3.4%
316800	criterion, max depth	"entropy", 1000	47%	3.1 %
316800	criterion, max depth, min samples split	"entropy", 1000 10	41%	8.7%
316800	criterion, max depth, min samples split	"entropy", 1000 15	39%	10.9%
316800	criterion, max depth, min samples split	"entropy", 1000 20	37%	12%
316800	criterion, max depth, min samples split	"entropy", 1000 1000	27%	22%
316800	criterion, max depth, min samples split, min samples leaf	"entropy", 1000 2000 1	26%	23%
316800	criterion, max depth, min samples split, min samples leaf, min weight fraction leaf	"entropy", 1000 2000 1 0.3	25%	24%

Table 1 is the result of accuracy obtained in training data and testing data with a total of 316800 data. It can be seen that to get good accuracy results, parameter changes are made when creating programs in python

from the use of criterion parameters, max depth and min samples split, min samples leaf and min weight fraction leaf. Then the decision tree parameter value is changed in order to get a good accuracy result. The more added the decision tree parameters, the more unstable and low the accuracy value. Therefore, the accuracy test is carried out on training data and testing data again as in table 2 below:

Table 2. Accuracy of the *decision tree*

Number of Datasets	Parameter *Decision Tree*	Parameter Value *Decision Tree*	Train Accuracy	Test Accuracy
79200	criterion, max depth, min samples split, min samples leaf, min weight fraction leaf	"entropy" 1000 2000 1 0.4	99%	99%

So for the results of the study, data as many as 316800 entered into the decision tree program turned out to be very low accuracy results. After checking again and conducting data training, it turns out that the data with the same input can cause different outputs. Therefore, when the machine learns it, the machine will be confused. Then the same input disposal is done but taken one only. So if the input is the same and there are many results, one will be taken only, namely the top of the order of the data. Then obtained from 316800 data to 79200 data with a good accuracy value of 99% and it turns out that the exact data is data for the severity is mild and not severe. So for those whose severity is moderate and severe it cannot be determined by this parameter. Because the image of the tree decision is a lot so to see, the results of the image can be viewed via the link below:
https://drive.google.com/drive/folders/1xtT3X4YRiwIxdiMCL_CYgLeMHFhw4TVT?usp=sharing

4. CONCLUSION

In making a classification model for people with the Covid-19 virus using machine learning methods with decision tree algorithms in Python programming can run as expected, although the training time carried out takes a long time. The study results indicate that the accuracy for the decision tree model in the dataset with the number of 316800 has a low accuracy value because data with the same input causes different outputs. Therefore, when the machine learns it, it will be confused. Then the disposal of the same input is done but only one is taken. So if there are the same inputs and there are many results, then only one will be taken, which is the top of the sequence of data. Then the accuracy rate of the results of data classification testing using decision trees is very high, with an accuracy of 99%. To get better results, when training on data depends on the length of training time and a lot of data tested. When testing the program requires sufficient and good laptop or computer specifications such as random access memory that is large enough for the program or when doing data training can be run properly.

REFERENCES

[1] Alvina Felicia Watratan, Arwini Puspita. B, and Dikwan Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.52158/jacost.v1i1.9.

[2] E. Supriatna, "Wabah Corona Virus Disease (Covid 19) Dalam Pandangan Islam," *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 6, 2020, doi: 10.15408/sjsbs.v7i6.15247.

[3] R. Rafiska, S. Defit, and G. W. Nurcahyo, "Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4.5," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 1, pp. 391–396, 2018, doi: 10.29207/resti.v2i1.275.

[4] A. Mujahidin and D. Pribadi, "Penerapan algoritma C4 . 5 untuk diagnosa penyakit pneumonia pada anak balita berbasis mobile," *J. Swabumi*, vol. 5, no. 2, pp. 155–161, 2017, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/swabumi/article/view/2523.

[5] M. A. Al Karomi, "Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm," *Jaict*, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.

[6] Y. Fitriani, S. Defit, and G. W. Nurcahyo, "Prediksi Hasil Belajar Siswa Secara Daring pada Masa Pandemi COVID-19 Menggunakan Metode C4.5," *J. Sistim Inf. dan Teknol.*, vol. 3, 2021, doi: 10.37034/jsisfotek.v3i3.149.

[7] R. Pakpahan and Y. Fitriani, "Analisa Pemafaatan Teknologi Informasi Dalam Pemeblajaran Jarak Jauh Di Tengah Pandemi Virus Corona Covid-19," *JISAMAR (Journal Inf. Syst. Applied, Manag. Account. Researh)*, vol. 4, no. 2, pp. 30–36, 2020.

[8] M. D. Ruhamak and E. H. Syai'dah, "Pengaruh Word Of Mouth, Minat Konsumen Dan Brand Image Terhadap Keputusan Konsumen (Studi Pada Pelajar Lembaga Kursus Di Area Kampung Inggris Pare)," *Ekonika J. Ekon. Univ. kadiri*, vol. 3, no. 2, p. 14, 2018, doi: 10.30737/ekonika.v3i2.186.

[9] Z. Zahrotunnimah, "Langkah Taktis Pemerintah Daerah Dalam Pencegahan Penyebaran Virus Corona Covid-19 di Indonesia," *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 3, pp. 247–260, 2020, doi: 10.15408/sjsbs.v7i3.15103.
[10] S. Syafrida and R. Hartati, "Bersama Melawan Virus Covid 19 di Indonesia," *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 6, pp. 495–508, 2020, doi: 10.15408/sjsbs.v7i6.15325.
[11] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018, doi: 10.33480/pilar.v14i2.926.
[12] A. Harun and D. P. Ananda, "Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decission Tree," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. April, pp. 58–63, 2021.
[13] D. T. Larose and C. D. Larose, "Discovering Knowledge in Data," *Discov. Knowl. Data*, 2014, doi: 10.1002/9781118874059.
[14] A. Ashari, "Paper_5-Performance_Comparison_between_Naïve_Bayes.pdf," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.
[15] D. R. Amancio *et al.*, "A systematic comparison of supervised classifiers," *PLoS One*, vol. 9, no. 4, 2014, doi: 10.1371/journal.pone.0094137.
[16] S. dan D. S. Sani, *Pengantar Data Mining : Menggali Pengetahuan dari Bongkahan Data*. Yogyakarta: Andi Offset. [Online]. Available: https://docplayer.info/53029387-_-pengantar-data-mining-menggali-pengetahuan-dari-bon-kahan-data-crh-3-t-o-o-t-ih9-t-7-l-ii-r_-j-_-i-p-r.html
[17] S. N. Hiadayat, A. R. I. Utami, and ..., "Penentuan Parameter Kinerja Bangunan Dengan Metode Inverse Modeling Menggunakan Machine Learning," *eProceedings …*, vol. 7, no. 1, pp. 1214–1220, 2020.
[18] F. D. Telaumbanua, P. Hulu, T. Z. Nadeak, R. R. Lumbantong, and A. Dharma, "Penggunaan Machine Learning," *J. Teknol. dan Ilmu Komput.*, vol. 3, no. 1, pp. 57–64, 2019.
[19] M. Batta, "Machine Learning Algorithms - A Review ," *Int. J. Sci. Res. (IJ*, vol. 9, no. 1, p. 381–undefined, 2020, doi: 10.21275/ART20203995.
[20] J. Eska, "Penerapan Data Mining Untuk Prediksi Penjualan Wallpaper Menggunakan Algoritma C4.5," vol. 2, 2018, doi: 10.31227/osf.io/x6svc.
[21] A. Krisna Ferdinan Leo Simanjuntak, Annita Carolina Br Barus, "Implementasi Metode Decision Tree Dan Algoritma C4.5 Untuk Klasifikasi Kepribadian Masyarakat," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 5, no. 1, pp. 51–59, 2021.
[22] G. D. M. Zulma and N. Chamidah, "Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree Dan K-Nearest Neighbor Pada Data Log Firewall," *Senamika*, no. April, pp. 679–688, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/view/1396

BIBLIOGRAPHY OF AUTHORS

Nadiah, currently as an active student of the final semester of Electrical Engineering, Diploma 4 Study Program of Telecommunication Engineering at State Polytechnic of Sriwijaya.

Sopian Soim, is currently as a lecturer in Electrical Engineering Diploma 3 and Diploma 4 Study Program of Telecommunication Engineering at State Polytechnic of Sriwijaya.

Sholihin, is currently as a lecturer in Electrical Engineering Diploma 3 and Diploma 4 Study Program of Telecommunication Engineering at State Polytechnic of Sriwijaya.