P-ISSN: 2655-8807
E-ISSN: 2656-8888

Application of the C4.5 Algorithm for Identifying Regional Zone Status Using A Decision Tree in the Covid-19 Series

Untung Rahardja[1]
Master of Information System[1]
University of Raharja, Tangerang Indonesia`
e-mail: untung@raharja.info

Author Notification
23 June 2022
Final Revised
18 July 2022
Published
21 July 2022

Rahardja, U. (2022). Application of the C4.5 Algorithm for Identifying Regional Zone Status Using a Decision Tree in the Covid-19 Series. Aptisi Transactions on Technopreneurship (ATT), 4(2), 164–173.

DOI: https://doi.org/10.34306/att.v4i2.234

Abstract

On July 25, 2020, there were 97,286 confirmed positive patients for COVID-19. It can be said that there was a pretty high increase in Indonesia. To identify patients quickly and accurately in a DKI Jakarta, revolutionary research is currently being carried out to assist medical personnel in simplifying their tasks. In this study, the C4.5 algorithm with the Rapid Miner software was used to identify the zone status of areas where the population is positive for COVID-19. The value possessed by this study can tie the mapping data of the level of the red, green, and yellow zones in which there are details about PDP, OPD, and other patient statuses. Mobile apps are the final results expected in this study to identify an area or zone of DKI Jakarta that is confirmed to be positive for COVID-19. This research shows that the desired results will make it easier for medical personnel to confirm the status of COVID-19 through the distribution of regional zone maps.

Keywords: *Covid-19, Application, Decision Tree, C4.5 Algorithm.*

1. Introduction

The COVID-19 was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and has since spread globally, resulting in a coronavirus pandemic [1]. The coronavirus pandemic was reported to have spread to Indonesia on 2 March 2020 [2]. The World Health Organization in April, through Covid-19 laboratory, confirmed 1.210.956 cases [3]. A fairly high increase was reported in positive patient cases every day in Indonesia. Many patients are positively confirmed Covid-19 in Jakarta. Then, Jakarta has become one of the coronavirus epicenters in Indonesia. It has become a special concern by the Central Government.

Departing from the case study above, the researchers researched the identification of the status of the zone during the pandemic corona, focusing on the current problem, which is affecting the capital city of Jakarta, Indonesia, and even almost in most countries around the world. In this study, the authors focus on the coronavirus pandemic problem or Covid-19 found in Indonesia specifically in Jakarta [4]. At this time there is no mobile application that can detect the status of zones in Jakarta, which identifies that this zone is a dangerous zone or red zone and can directly provide detailed information on the number of positive patients, patients under surveillance, people under surveillance, number of treated, the number who died, the number who recovered when we were in the position of the area [5].

We use the C4.5 algorithm to design the system modeling to identify the status of the corona zone in Jakarta [6]. The author uses data mining techniques for data processing. According to the author, the C.45 algorithm is the right one used in this study. Because in the C4.5 algorithm, sample data and training data can be made input that can produce decision tree output [7]. The data from the decision tree produces data that is easy to get information and knowledge [8].

2. Related Work

Data mining is a scientific discipline that studies methods for extracting knowledge or finding patterns from data [9]. And the results of data processing can be used to make decisions in the future [10], used to process data, and extract big data so that it becomes valuable by using new knowledge [11]. In general, data mining steps are starting from setting goals, data preparation, data preprocessing, data mining, and knowledge evaluation [12].

The process of improving communication between members continuously is essential to achieving good and right goals [13]. But it is a difficult process, if this process cannot be done properly it will affect the next process. Therefore determining goals is important to produce data mining output which is a good decision based on discovery and knowledge [14].

There is some previous research that is relevant to our study. They were, research for determining the types of road construction using analytic hierarchy process [15], diagnosing chronic kidney disease using C4.5 algorithm [16], design and construction on navigation [17], and decision support system model using C4.5 algorithm [18]. But this research has not discussed in depth about how to prevent spread Covid-19 in contextualized regional zones using computer-based application systems.

3. Research Methods

The research method used in this study consisted of five stages. It consists of study literature, data collection, data processing and validation, testing and analysis, and evaluation. The research method is declared in detail as the research design.

3.1 Research Design

The first phase of the research design in this study is study literature. It is to find out various previous studies on Covid-19, which have links with research conducted. Perform identification and formulation of problems and determine research objectives [4]. The second phase is the data collection. In the phase data collection, we focused on collecting data consisting of primary data by downloading data from the Covid-19 website and Jakarta corona website regarding the data collection map of corona patient villages throughout Jakarta [19].

In the data preparation process, this must also be really prepared, to avoid lack of data, data loss, data cannot be found or incomplete data. It was not easy and required hard work in the beginning before the data mining process [20]. After the data preparation phase is complete, then the preprocessing process is a classification or grouping of a raw data model, in a simpler and more effective form used for the data mining process [21].

The third phase is data processing and validation. The primary data processing using Rapid Miner5 software and validating the data that has been processed. Then the fourth phase is system design. In the fourth phase, we perform system design using data mining and C4.5 algorithm to form a new system [20]. And the fifth phase, testing and analysis. In the fifth phase to ensure the system meets expectations according to the requirements specifications. Finally, we made an evaluation to the final system.

3.2 Decision Tree and C4.5 Algorithm

The C4.5 algorithm was introduced by Quinlan to induce a classification model and is also called a decision tree whose data is based on training data provided by getting rations [17]. Classification is a data mining technique that can be used to predict group membership to data instances [21]. And C4.5 Decision Tree is the first supervised fundamental machine learning classification algorithm that is widely applied and usually achieves excellent performance in predictions [22]. A decision tree is a tree structure like a flowchart that can be constructed from a given set of attributes, where each branch represents a test result, and each leaf node represents a class [23].

The concepts in the decision tree, data is expressed in tabular form with attributes and records, and attribute states a parameter created as a criterion in the formation of a tree. One attribute is an attribute that states the per-item data solution called the target attribute, and attributes have values named instance [23] [24] [25].

On the other hand, C4.5 algorithm is an extension of ID3. The speed of C4.5 is significantly faster than ID3 (faster in some order of magnitude) and C4.5 is more memory efficient than ID3. The resulting decision tree is the result of the C4.5 algorithm and can represent and model significant data exploration results so that the knowledge or information from this data is more easily identified [26]. Some development that has been done for C4.5 can overcome missing values, overcome advanced data, and pruning.

In general, using the C4.5 algorithm to build a decision tree is: select the attribute as root, create branches for each root, divide cases into branches, and repeat the process for each branch until all cases in the branch have the same class. To choose the attribute as root, based on the highest gain value of the existing attribute. To calculate the gain, use the formula in the equation below [26]:

$$\text{Gain (S, A)} = \sum_{i=1}^{n} \frac{|si|}{|s|} Entropy\ (Si) \quad (1)$$

With: S as a set of cases; A as an attribute; n as the number of partitions in attribute A; | Si | as the number of partitions i; | S | as a set of cases in S. Meanwhile, the entropy value can be seen in equation 2:

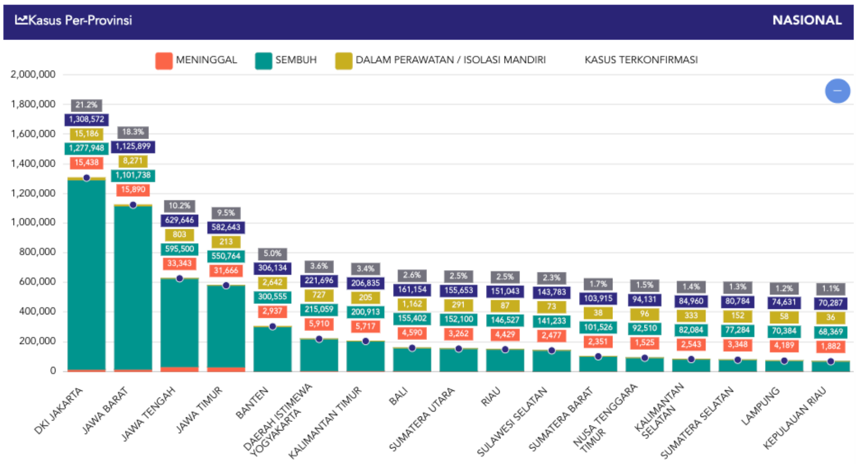
$$\text{Entropy (Si)} = \sum_{i=1}^{n} -\ pi \log log\ 2\ pi \quad (2)$$

With: S as a set of cases; A as a feature; n as the number of S partitions; Pi as proportion of S i to S. There are 3 types of nodes in the decision tree, as follows, root node, internal node, and leaf node [26]. The first is the root node, which is the top node, this node has no input and cannot have any output or has more than one output. The second is an internal node, which is a branching node, this node can only have one input and two minimal outputs. And the third is the Leaf node or terminal node, which is the last node, this node can only have one input and has no output [27].

4. Conclusion

Based on observations, there are currently no applications that can directly identify areas or zones that have been identified as positive corona or red or green status zones and can display corona patient data in that region. From the analysis and observation above, the writing makes application design that can identify a city, especially Jakarta.

Application in this research provided information about the status of the zone in the area or city is red or green, the number of ODP, the number of PDP, the number of positive corona patients, as well as the number of treated, cured and died using data mining methods and C4.5 algorithm to determine the status of the zone in the city [27]. The data used in this study is the data that the writer took from the corona DKI website [28], namely Covid-19 positive distribution map data. Based on the data obtained by the authors, for example, there are 267 data spreads of Covid-19 positive cases19.

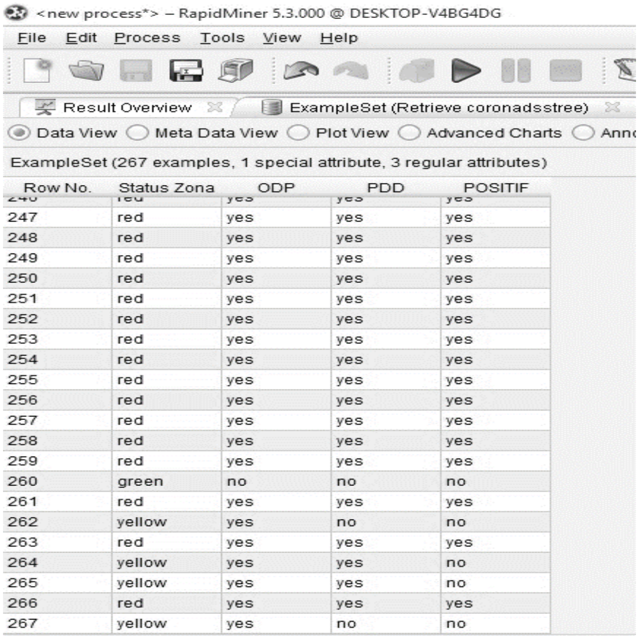


Picture 1. Data Distribution of Positive Case of Covid19 DKI Province

The table above is a data distribution of positive cases of Covid-19 in Jakarta. In the ODP column, there is ODP (insider oversight) data, PDP column is the data of patient under supervision status, the positive column is the data of positive patient status, and the last column, the regional zone status of Covid-19. Data obtained by the author. The author thought by using data mining. Data mining is the process of extracting between the lines, not previously known but potentially useful information and knowledge from large amounts of incomplete, noisy, unclear, random data [8] [29]. From a large amount of data in the positive case report table Covid-19 in Jakarta. The authors extract the data to make a decision about the status of the zone in the area affected by the Covid-19 pandemic by using the 4.5 algorithms as a decision tree.

ODP (insider oversight) data, PDP (patient under supervision), and positive patient data are the main data used in data processing in the data mining and decision tree processes using the C4.5 algorithm. In this research, we use 10 sample data out of 267 of the data. The three components in the ODP, PDP, and Positive columns can determine the status of the zone whether Red, Yellow, or Green. The algorithm to determine the region or city is a red, yellow, or green zone with data mining using the C4.5 algorithm in which there is a decision tree method.

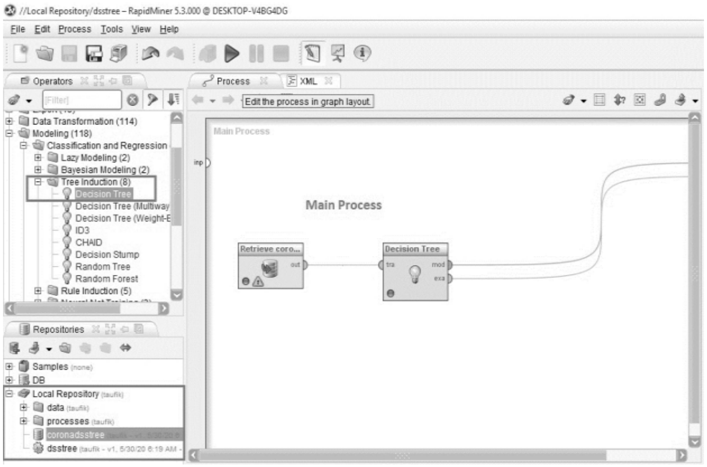
The database of the corona positive patient distribution data is processed using rapid miners to produce a decision tree. Sample data from 267 cases of ODP, PDP, and Covid-19 positive imported into the RapidMiner database. The data that is imported into rapid miners shown in Figure 1.And data is processed using decision tree data modeling using Rapid Miner shown in Picture 2 and Picture 3.


<new process*> – RapidMiner 5.3.000 @ DESKTOP-V4BG4DG
File Edit Process Tools View Help
Result Overview
ExampleSet (Retrieve coronadsstree)
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Ann

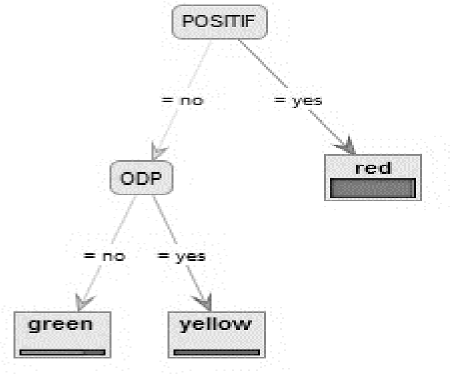
ExampleSet (267 examples, 1 special attribute, 3 regular attributes)

Row No.	Status Zona	ODP	PDD	POSITIF
246	red	yes	yes	yes
247	red	yes	yes	yes
248	red	yes	yes	yes
249	red	yes	yes	yes
250	red	yes	yes	yes
251	red	yes	yes	yes
252	red	yes	yes	yes
253	red	yes	yes	yes
254	red	yes	yes	yes
255	red	yes	yes	yes
256	red	yes	yes	yes
257	red	yes	yes	yes
258	red	yes	yes	yes
259	red	yes	yes	yes
260	green	no	no	no
261	red	yes	yes	yes
262	yellow	yes	no	no
263	red	yes	yes	yes
264	yellow	yes	yes	no
265	yellow	yes	yes	no
266	red	yes	yes	yes
267	yellow	yes	no	no

Picture 2. Sample Data Positive CaseCOVID19



Picture 3. Model analysis using Rapid Miner



Picture 4. Decision Tree Using Rapidminer

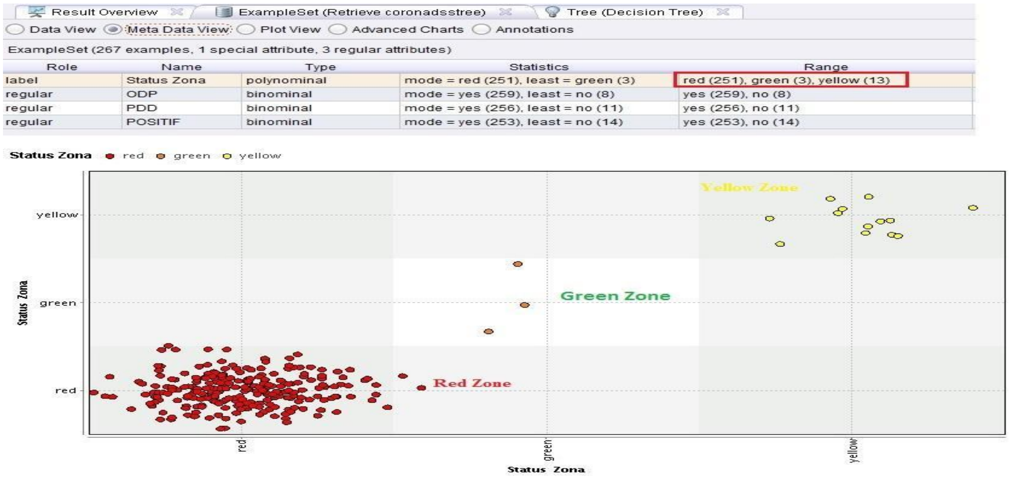
The results of processing by Rapid Miner produces a decision tree like the Figure 3. And produced three rule bases, they are: if there are positive patients then the red zone. If there is no positive but there is an ODP or PDP patient then the yellow zone, and if there are no positive patients and there is no ODP or PDP then the green zone status.

In this research, we developed an application to determine the status of the regional zone and prevent the spread of Covid-19in Indonesia using the data in Table 1. It also integrated with a Geographic Information System and Google Maps API and GPS systems. The advantage of the API is that it allows an application with other applications to interact and interact. These features make Google Maps JavaScript API, the most commonly used Maps API for online mapping. A picture of the information system architecture to determine the status of the regional zone during the Covid-19 pandemic shown in Picture 4.

This information system which is proposed in this research has some strengths and weaknesses. The strengths are, it can provide accurate and real-time information about the status of the corona zone according to the coordinates, and provide information on ODP, PDP, and corona positive patients. It can be further developed because the system is designed with Google Maps API integration, such as the Covid-19 human on pandemic tracking system [30]. But, it has some weaknesses, a large enough server with high and stable specifications must be prepared, a stable supply of electricity, and UPS is needed for 24 hours, takes a fast and stable server-side internet, and safe security is needed.

But also, this system in this study can be used as an application of decision systems when in the Corona zone to maintain personal and family health safety. So, this result has a function which is a little similar to the other application about Covid-19 [31].

Based on the results of data processing in the database using the C4.5 algorithm in Figure 5.It showed that the data in the database and the data that appears on the map showing the red zone, green zone, and yellow zone. The data in the database that puts a checkbox in Picture 5. It shows the red zone appears to be 251, in the green zone 3 and the yellow zone is 13. And the user can look at the map the number of nodes that appear according to the nodes in the green zone, yellow and red.


Result Overview
ExampleSet (Retrieve coronadsstree)
Tree (Decision Tree)
○ Data View ● Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (267 examples, 1 special attribute, 3 regular attributes)

Role	Name	Type	Statistics	Range
label	Status Zona	polynominal	mode = red (251), least = green (3)	red (251), green (3), yellow (13)
regular	ODP	binominal	mode = yes (259), least = no (8)	yes (259), no (8)
regular	PDD	binominal	mode = yes (256), least = no (11)	yes (256), no (11)
regular	POSITIF	binominal	mode = yes (253), least = no (14)	yes (253), no (14)

Status Zona ● red ● green ● yellow
Yellow Zone
Green Zone
Red Zone
yellow
green
red
Status Zona
red
green
yellow
Status Zona

Picture 6. Regional zone of Covid-19

5. Discussion

Figure 5 shows the status of Covid-19 in Jakarta [32]. The status are red zone, yellow zone, and green zone. Based on the status region, the government must carry out large-scale social restrictions for the red zone. For all residents in the red zone area who want to carry out activities required to use personal protective equipment that is required to use a mask, prepare a place to wash their hands with soap or hand sanitizer every time they enter the workplace or into the home. Changing clothes after activity from home and must immediately go for a bath. The local government is required to conduct rapid tests on residents who have identified symptoms of Covid-19to reduce the spread. These results complete the results of other studies about Covid-19 [31].

For areas that are in the yellow zone, safety must be improved for health, that is, the government must implement regulations so that citizens do not enter the red zone, use personal protective equipment or masks for every activity, wash their hands with soap or hand sanitizer after every activity [28]. Change clothes or clothes and wash yourself with a shower after finishing work. For areas that are green zones, the local government must continue to maintain a healthy lifestyle and always maintain hygiene that has resulted in areas that have been protected from positive cases, PDP, or ODP [33].

And for areas with green zones, the government can be Raw Models for areas that are still yellow or red zones or zones with green zone status. Then it can socialize the culture of healthy living within the environment of residence or in daily life via communication media or media other technologies such as whatsapp group, Facebook, IG, or others [34].

6. Conclusion

This research has proposed an application for determining the status of the regional zone of Covid-19 in Jakarta. An Application developed using Decision Tree and C4.5 algorithms to analyze the data Covid-19.The application shows that there are three zones which confirm Covid-19 in Jakarta. They are red, yellow and green. The red zone is the most dominant in Jakarta. We suggest further development applications to identify an outbreak of the disease in a city or even every province in Indonesia. And the information system can be developed by integrating with the national population.

References

[1] Y. Jia and X. Wang, "Intelligent Traffic Decision Analysis System Based on Big Data Mining," MS&E, vol. 392, no. 6, p. 62187, 2018.

[2] N. Hatami et al., "Worldwide ACE (I/D) polymorphism may affect COVID-19 recovery rate: an ecological meta-regression." Springer, 2020.

[3] A. Husnayain, A. Fuad, and E. C.-Y. Su, "Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan," Int. J. Infect. Dis., 2020.

[4] U. Rahardja, S. Sudaryono, N. P. L. Santoso, A. Faturahman, and Q. Aini, "Covid-19: Digital Signature Impact on Higher Education Motivation Performance," Int. J. Artif. Intell. Res., vol. 4, no. 1, May 2020, doi: 10.29099/ijair.v4i1.171.

[5] Q. Aini, S. Riza Bob, N. P. L. Santoso, A. Faturahman, and U. Rahardja, "Digitalization of Smart Student Assessment Quality in Era 4.0," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1.2, pp. 257–265, Apr. 2020, doi: 10.30534/ijatcse/2020/3891.22020.

[6] U. Rahardja, A. N. Hidayanto, T. Hariguna, and Q. Aini, "Design Framework on Tertiary Education System in Indonesia Using Blockchain Technology," 2019 7th Int. Conf. Cyber IT Serv. Manag. CITSM 2019, pp. 5–8, 2019, doi: 10.1109/CITSM47753.2019.8965380.

[7] D. M. B. Tarigan and D. P. Rini, "Particle Swarm Optimization–Based on Decision Tree of C4. 5 Algorithm for Upper Respiratory Tract Infections (URTI) Prediction," in Journal of Physics: Conference Series, 2019, vol. 1196, no. 1, p. 12077.

[8] I. H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations,” Acm Sigmod Rec., vol. 31, no. 1, pp. 76–77, 2002.

[9] A. Asyary and M. Veruswati, “Sunlight exposure increased Covid-19 recovery rates: A study in the central pandemic area of Indonesia,” Sci. Total Environ., p. 139016, 2020.

[10] A. Pradipta, D. Hartama, A. Wanto, S. Saifullah, and J. Jalaluddin, “The Application of Data Mining in Determining Timely Graduation Using the C45 Algorithm,” IJISTECH (International J. Inf. Syst. Technol., vol. 3, no. 1, pp. 31–36, 2019.

[11] E. Indra, K. Ho, R. Hakim, D. Sitanggang, and O. Sihombing, “Application of C4. 5 Algorithm for Cattle Disease Classification,” in Journal of Physics: Conference Series, 2019, vol. 1230, no. 1, p. 12070.

[12] R. Zhuo and Z. Bai, “Key technologies of cloud computing-based IoT data mining,” Int. J. Comput. Appl., pp. 1–8, 2020.

[13] U. Rahardja, E. P. Harahap, and S. R. Dewi, “The strategy of enhancing article citation and H-index on SINTA to improve tertiary reputation,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 17, no. 2, pp. 683–692, 2019, doi: 10.12928/TELKOMNIKA.V17I2.9761.

[14] Sudaryono, U. Rahardja, and Masaeni, “Decision Support System for Ranking of Students in Learning Management System (LMS) Activities using Analytical Hierarchy Process (AHP) Method,” J. Phys. Conf. Ser., vol. 1477, no. 2, 2020, doi: 10.1088/1742-6596/1477/2/022022.

[15] H. Henderi, E. Kurnadi, and D. Trisnawarman, “Decision Support System Model Determines the Type of Road Construction in Indonesia,” in IOP Conference Series: Materials Science and Engineering, 2020, vol. 852, no. 1, p. 12142.

[16] M. A. Muslim, A. J. Herowati, E. Sugiharti, and B. Prasetiyo, “Application of the pessimistic pruning to increase the accuracy of C4. 5 algorithm in diagnosing chronic kidney disease,” in Journal of Physics: Conference Series, 2018, vol. 983, no. 1.

[17] M. T. H. Nyo and W. Z. Hein, “Design and Construction of Navigation Based Auto Self-Driving Vehicle using Google Map API with GPS,” Int. J. Trend Sci. Res. Dev, vol. 3, pp. 65–68, 2019.

[18] C. E. A. Pah and D. N. Utama, “Decision Support Model for Employee Recruitment Using Data Mining Classification,” Int. J., vol. 8, no. 5, 2020.

[19] T. Hariguna, E. P. Harahap, and Salsabila, “Implementation of Business Intelligence Using Highlights in the YII Framework based Attendance Assessment System,” Aptisi Trans. Technopreneursh., vol. 1, no. 2, 2019, doi: 10.34306/att.v1i2.32.

[20] U. Rahardja, T. Hariguna, and W. M. Baihaqi, “Opinion mining on e-commerce data using sentiment analysis and k-medoid clustering,” Proc. - 2019 12th Int. Conf. Ubi-Media Comput. Ubi-Media 2019, pp. 168–170, 2019, doi: 10.1109/Ubi-Media.2019.00040.

[21] S. Shokouhyar, P. Saeidpour, and A. Otarkhani, “Predicting Customers’ Churn Using Data Mining Technique and its Effect on the Development of Marketing Applications in Value-Added Services in Telecom Industry,” Int. J. Inf. Syst. Serv. Sect., vol. 10, no. 4, pp. 59–72, 2018.

[22] U. Rahardja, C. Lukita, F. Andriyani, and Masaeni, “Optimization of marketing

workforce scheduling using metaheuristic genetic algorithms," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1.2 Special Issue, pp. 243–249, 2020, doi: 10.30534/IJATCSE/2020/3691.22020.

[23] N. Priyanka and P. RaviKumar, "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree," in 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2017, pp. 1–7.

[24] S. Nayak, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "Prediction of Heart Disease by Mining Frequent Items and Classification Techniques," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 607–611.

[25] H. W. Ian and F. Eibe, "Data mining: Practical machine learning tools and techniques." Morgan Kaufmann Publishers, 2005.

[26] H. Xu, "The study on eco-environmental issue of Aral Sea from the perspective of sustainable development of Silk Road Economic Belt," in IOP Conference Series: Earth and Environmental Science, 2017, vol. 57, no. 1, p. 12060.

[27] P. A. Sunarya, F. Andriyani, Henderi, and U. Rahardja, "Algorithm automaticPrawira, M., Sukmana, H. T., Amrizal, V., & Rahardja, U. (2019). A Prototype of Android-Based Emergency Management Application. 2019 7th International Conference on Cyber and IT Service Management, CITSM 2019. https://doi.org/10.1109/CI," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 1.5 Special Issue, pp. 387–391, 2019, doi: 10.30534/ijatcse/2019/6281.52019.

[28] Q. Aini, U. Rahardja, I. Handayani, M. Hardini, and A. Ali, "Utilization of google spreadsheets as activity information media at the official site alphabet incubator," Proc. Int. Conf. Ind. Eng. Oper. Manag., no. 7, pp. 1330–1341, 2019.

[29] Y. Liu et al., "A COVID-19 Risk Assessment Decision Support System for General Practitioners: Design and Development Study," J. Med. Internet Res., vol. 22, no. 6, p. e19786, 2020.

[30] S. L. Pan and S. Zhang, "From fighting COVID-19 pandemic to tackling sustainable development goals: An opportunity for responsible information systems research," Int. J. Inf. Manage., p. 102196, 2020.

[31] D. S. Hui et al., "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China," Int. J. Infect. Dis., vol. 91, pp. 264–266, 2020.

[32] R. Tosepu, J. Gunawan, D. S. Effendy, H. Lestari, H. Bahar, and P. Asfian, "Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia," Sci. Total Environ., p. 138436, 2020.

[33] R. Bakhtiar, H. Hilda, K. Duma, and R. C. P. Yudia, "Relationship between understanding of COVID-19's infographics and the efforts to prevent COVID-19 transmission," J. Community Empower. Heal., vol. 3, no. 2.

[34] T. Hariguna, U. Rahardja, Q. Aini, and Nurfaizah, "Effect of social media activities to determinants public participate intention of e-government," Procedia Comput. Sci., vol. 161, pp. 233–241, 2019, doi: 10.1016/j.procs.2019.11.119.