



Implementation of Data Analytics for the Accuracy of Service Time Prediction Models

Ritzkal Ritzkal*, Fitriah Rachmawati, Dahlia Widhyaestoeti, Fety Fatimah

Teknik Informatika, Universitas Ibn Khaldun, Jl. Kh. Sholeh Iskandar Km 2, Bogor 16162, Indonesia

Corresponding Author Email: ritzkal@ft.uika-bogor.ac.id

<https://doi.org/10.18280/mmep.090633>

ABSTRACT

Received: 4 June 2022

Accepted: 22 October 2022

Keywords:

accuracy, decision tree J48, classification, prediction, weka

The performance of repair services is very important to determine the achievement of consumer confidence. The strategy that needs to be made is to pay attention to the timeliness of repairs so that no one is harmed between consumers and service providers. Tracing information on data from repair services is one effective way to determine the accuracy of computer repair time. The collection of information used comes from the Istdata Indopacific Solution Center (IISC) repair service dataset, consisting of a collection of data on the completion time of product unit repairs that are achieved and not achieved. Repair completion time is the time in accordance with the agreement between the repair service party and the consumer. Data processing is carried out by processing analytical data by utilizing the Weka Tools software with the application of classification with the J48 decision tree method which is the development of the C4.5 algorithm. The effectiveness of this method was tested using 10-fold cross validation, where from the results of the confusion matrix measurement an accuracy of 99.5% was obtained. The result states that the J48 decision tree method is effective and can be used to predict the accuracy of computer repair time.

1. INTRODUCTION

Currently, there are computer repair services available that cater to individual consumers and companies who need computer repair services. In a day, data from computer repair services can receive many units of computer damage. To repair the damage, it takes a completion time in accordance with the specified time.

When the repair is completed, an agreement on the length of the repair is made (Service Level Agreement) with the consumer so that there is no misunderstanding. Even though the agreement and completion time have been determined, sometimes there are obstacles in the repair process, such as complexity in analysis, inappropriate comparisons, and spare part stock that is not yet available in the warehouse.

Time of repair needs to be considered for accuracy so that no one is harmed between consumers and service providers of repair services, both morally and materially. Data analytic techniques are one of the effective ways to predict the timeliness of repairs by integrating heterogeneous data from various sources which are then drawn into a conclusion and made into a prediction for strategic decision making [1]. In data analytics, there is a data mining process. Data mining is a specific process to find scientific and observable knowledge, such as patterns, association relationships, changes, peculiarities of data and the structure of data in databases and information storage [2]. Meanwhile, Decision trees have an important role in producing science, business, and engineering in various fields, including classes on the use of data mining techniques that are easy to understand, easy to use and produce excellent decision results [3].

J48 decision tree method produces an effective level of accuracy for predicting diabetes mellitus diagnosis [4].

The level of accuracy using the J48 decision tree [5] can be seen from the results of the confusion matrix calculation resulting from the weka data mining application [6]. Where the data in the matrix is the result of the evaluation of the performance of the weka data mining application [7]. Based on the explanation of the above phenomena, problems can be identified, namely: (1). Determination of the repair time from the agreed one is not on time due to the complexity of the analysis, inappropriate comparisons and the unavailability of spare part stock in the warehouse. (2). The inaccuracy of repair time results in losses between consumers and computer repair service providers, both morally and materially. From the formulation of the problem above, it can be concluded as follows: (a) Problem Statement The timing of computer repairs is still low, so there is no time prediction for the completion of computer repairs. (b) Research Question. The research question that can be asked in this study is "How does the application of the J48 decision tree method determine the accuracy and predictive model of computer repair time correctly?". As for the aims and objectives of this research are as follows: 1). Meaning. The purpose of this research is to apply the J48 decision tree method to determine the accuracy of computer repair time. 2). Purpose. The aim of this research is to determine the level of accuracy of the J48 decision tree method in determining the accuracy of computer repair time.

2. METHOD

This research starts from the problem identification stage, this stage is carried out after obtaining the appropriate dataset to be carried out at the classification stage. At this stage, the data that has been collected is pre-processed by changing the

type of file.xls to file.csv. The data pre-processing stage includes three stages of the process, namely the data selection stage, data cleaning, and the attribute discretization stage. Furthermore, the process of compiling the J48 decision tree uses the weka data mining application.

The data in this study were collected by conducting observations and interviews. Observation is done by observing to get the required data. To support the observations in finding out information about conditions related to the repair time and possible obstacles that occurred during the repair process, then conducted an interview process. The data that has been collected is analyzed with a confusion matrix. The data obtained from the confusion matrix is calculated using the formula for prediction accuracy and error rate to get the results in the form of percentage values. The results of this analysis are used to determine the effectiveness of the method. Accuracy is used as an assessment measure. If a profit matrix is available, then profitability can be used as an assessment measure. Accuracy can be calculated based on a training sample, a data validation set, and a cross-validation approach [3].

3. RESULT

3.1 Data collection

The data carried out by the data mining process in this study were from the service center service dataset. The amount of data obtained is in the form of reports per day within a grace period of six months from January to June as many as 5,312 records with 15 attributes. There are several basic attributes for grouping Service Level Agreement (SLA) [8] achieved and Service Level Agreement (SLA) not achieved, namely [9]:

1. Product Line

It is an attribute that contains the type- product type.

2. Case Type

It is an attribute that contains a description of the status of the product that is still under warranty or is out of warranty.

3. Turn Around Time (TAT)

Is an attribute that contains a grace period for repairing product damage in days.

4. Service Level Agreement (SLA)

Is an attribute that contains a description of the achievement of the duration of the repair according to the agreement with the consumer is achieved or not achieved.

3.2 Problem identification

1) Data Pre- Processing Stage

The data pre-processing stage is the beginning of the data mining process [10]. Data pre-processing includes identification and selection of attributes, checking for inconsistent data, namely handling incomplete attribute values (missing values), and attribute discretization process. Identification and selection of attributes is an initial requirement for the data mining process which will result in the presence or absence of complete values for each attribute that will be used at the data mining stage. Incomplete attribute (missing value) if there is a certain record in one of the attributes of the missing record value, the record in question will be deleted, because the record is considered to be missing data or missing value. The next stage is to discretize attributes to make it easier to group values based on predetermined

criteria. It aims to simplify the problem simplification and increase accuracy in the process.

Product line attributes are divided into two groups, namely mobility and peripherals. Case type attributes are divided into two groups, namely warranty and out of warranty (oow). The Turn Around Time (TAT) attribute is divided based on the numeric data type of the total number of repairs carried out in a matter of days. The Service Level Agreement (SLA) attributes are divided into 2 groups, namely achieved or not achieved. The next stage is the transformation of the file type so that it can be read by the weka data mining application. The results of the transformation can be seen as in Table 1.

2) J48 Decision Tree Preparation

At this stage, the data is implemented with the C4.5 algorithm formula which produces J48 in the weka classifier package to build a tree [11]. To select an attribute as root, based on the highest value of the existing attributes, Entropy and Information Gain are used. The stage begins with doing, taking sample data or samples from data resolved and canceled cases. Where the data can be seen in Table 2.

The gain value is obtained for each attribute, where the attribute with the highest gain value will be the parent for the next nodes. These nodes come from attributes that have a gain value that is smaller than the gain value of the parent attribute. So to get the gain value from two different output classes, namely Service Level Agreement (SLA) was achieved (achieved) and Service Level Agreement (SLA) was not achieved (not achieved) in the dataset resolved and canceled cases. The steps in forming a decision tree are counting the number of cases.

Table 1. Data resolved and cancelled cases (DATAR&C).csv

Product	CaseType	TAT	SLA
Mobility	Oow	4	Achieved
Mobility	Oow	6	Achieved
Mobility	warranty	4	Not
Mobility	Oow	36	Not
Mobility	Oow	4	Achieved
Mobility	warranty	4	Not
Mobility	Oow	2	Achieved
Mobility	Oow	6	Achieved
Mobility	Oow	33	Not
Mobility	Oow	0	Achieved
Mobility	Oow	7	Achieved
Mobility	warranty	0	Achieved
Mobility	warranty	4	Not
Mobility	warranty	0	Achieved
Mobility	Oow	6	Achieved
Mobility	Oow	0	Achieved
Mobility	warranty	0	Achieved
Mobility	warranty	0	Achieved
Peripheral	Oow	5	Achieved
Mobility	warranty	4	Achieved
Mobility	warranty	4	Achieved
Mobility	warranty	4	Achieved
Mobility	Oow	0	Achieved
Mobility	warranty	4	Not
Mobility	Oow	4	Achieved
Mobility	Oow	0	Achieved
Peripheral	Oow	0	Achieved

The number of cases for the Service Level Agreement (SLA) decision was achieved (achieved), the number of cases for the Service Level Agreement (SLA) decision was not achieved (not achieved), and the entropy of all cases and cases divided based on product line attributes, case type, Turn

Around Time (TAT). After that, calculate the gain of each attribute for determining the root node. The results of the calculation of the gain of each attribute for determining the root node can be seen in Table 3.

Row Total Entropy column in Table 3 is calculated by the following equation: $Entropy(S) = \sum_{i=1}^n - pi \times \log_2 pi$

$$Entropy(Total) = - \left(\left(\frac{4054}{5310} \right) \times \log_2 \left(\frac{4054}{5310} \right) \right) - \left(\left(\frac{1256}{5310} \right) \times \log_2 \left(\frac{1256}{5310} \right) \right)$$

$$Entropy(Total) = 0,789230196$$

Mean while, the gain value on the product line line is calculated using the following equation:

$$Gain(S, A) =$$

$$Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i)$$

$$Gain(Total, prod) =$$

$$0,789230196 - \left(\left(\frac{4898}{5310} \right) \times 0,77 \right)$$

$$- \left(\left(\frac{1256}{5310} \right) \times 0,88 \right)$$

$$Gain(Total, prod) = 0,15$$

Table 2. Dataset after pre-process

Product	CaseType	TAT	SLA
Mobility	Oow	4	Achieved
Mobility	Oow	6	Achieved
Mobility	warranty	4	Not
Mobility	Oow	36	Not
Mobility	Oow	4	Achieved
Mobility	warranty	4	Not
Mobility	Oow	2	Achieved
Mobility	Oow	6	Achieved
Mobility	Oow	33	Not
Mobility	Oow	0	Achieved
Mobility	Oow	7	Achieved
Mobility	warranty	0	Achieved
Mobility	warranty	4	Not
Mobility	warranty	0	Achieved
Mobility	Oow	6	Achieved
Mobility	Oow	0	Achieved
Mobility	warranty	0	Achieved
Mobility	warranty	0	Achieved
Peripheral	Oow	5	Achieved
Mobility	warranty	4	Achieved
Mobility	warranty	4	Achieved
Mobility	warranty	4	Achieved
Mobility	Oow	0	Achieved
Mobility	warranty	4	Not
Mobility	Oow	4	Achieved
Mobility	Oow	0	Achieved
Peripheral	Oow	0	Achieved

Table 3. Node calculations

Node	Amount Case (S)	Ach (S1)	Not (S2)	Entropy	Gain
1 Total Prod.	5310	4064	1256	0,789230196	0,1475635
Mob.	4897	3791	1106	0,770711028	
Perip.	413	288	125	0,884522339	
C. Type					0,6405824
Warr.	3289	2587	702	0,748004545	
Oow	2021	1482	529	0,829378087	
TAT					0,7892302
<=3	3661	3661	0	0	
>3	336	0	366	0	
<=7	393	393	0	0	
>7	890	0	890	0	

From the results in the table, it can be seen that the attribute with the highest gain is Turn Around Time (TAT), which is 0.79. Thus, Turn Around Time (TAT) can be a root node. Where, there are four attribute values of Turn Around Time (TAT), namely TAT less than equal to three (<=3), more than three (>3), less than equal to seven (<=7) and more than seven (>7). Of the four values, attribute values less than equal to three (<=3) and less than equal to seven (<=7) have classified cases into 1, namely the decision was achieved (achieved), and for attribute values more than three (>3) and more than seven (>7) have classified cases into 0, namely the decision was not achieved (not achieved) so that no further calculations are needed. From these results it can be described that the temporary decision tree looks like in Figure 1.

3.3 Implementation of J48 using weka 3.7.9

Data resolved and canceled cases (DATAR&C).csv processed using weka tools with the application of decision tree J48 which is the development of the C4.5 algorithm to determine the grouping of Service Level Agreement (SLA)

achieved (achieved) or Service Level Agreement (SLA) not achieved (not achieved).

1) Pre-processing stages on weka 3.7.9

In the first stage weka is a pre-process, namely by entering the data resolved and canceled cases (DATAR&C).csv as the main dataset to be classified. The total dataset resolved&cancelled cases consisted of 5310 records and 4 attributes. Weka will explore the characteristics of the attributes of the dataset. The pre-process stages in Weka 3.7.9 can be seen in Figure 2 and Figure 3.

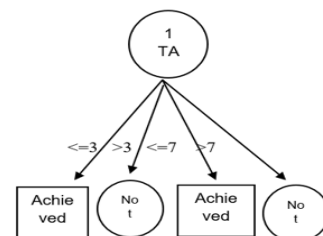


Figure 1. Decision tree calculation results node

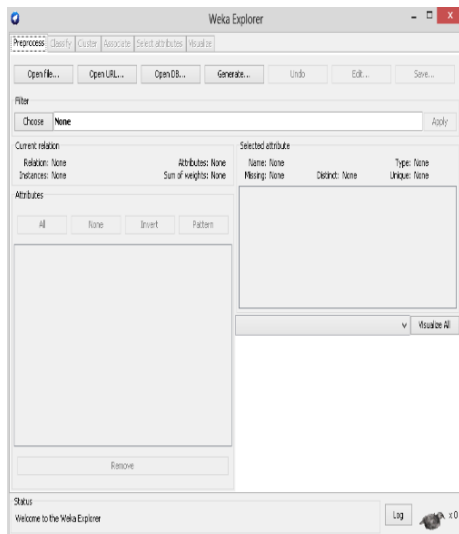


Figure 2. Display of weka pre-process 3.7.9 before entering data

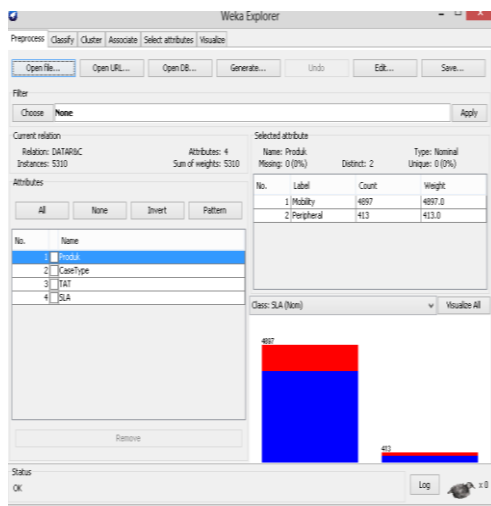


Figure 3. Display of weka pre-process 3.7.9 after entering data

The dataset resolved and canceled cases were processed using the J48 classifier technique with the output of a Service Level Agreement (SLA). The type of test used is cross validation [12]. The display of the classify panel on weka can be seen in Figure 4.

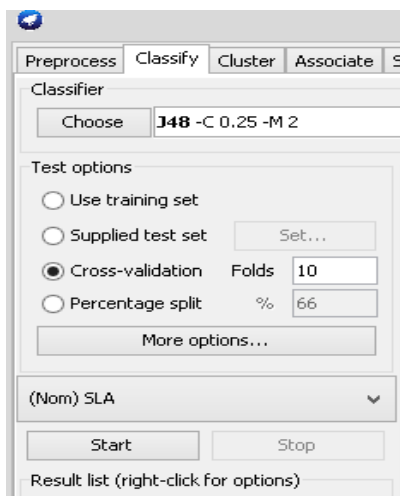


Figure 4. Panel classifier information

1. *Choose*: Selection of the classification algorithm to be used.
2. *Use Training Set*: Using training data sets.
3. *Supplied test set*: Using data testing.
4. *Cross-validation*: Data sharing.
5. *Percentage Split*: Percentage of splits or ramifications.

After doing the classifier technique from the dataset, and selecting the type of test used and then clicking start, the classifier output results come out. As in Figure 5.

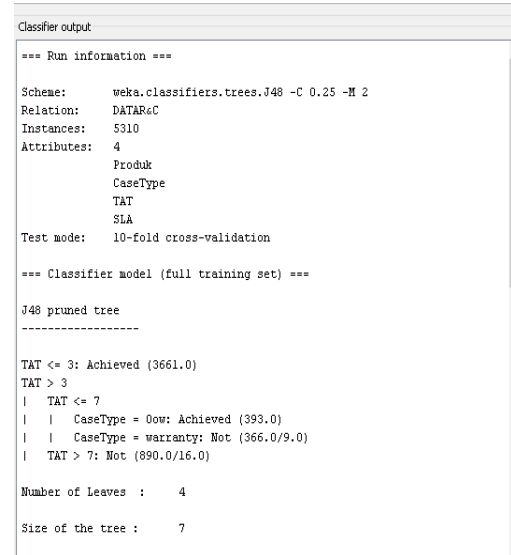


Figure 5. Classifier output

The classification process [13] is influenced by the selected attributes that support to determine the service level group / Service Level Agreement (SLA) is achieved (achieved) and the Service Level Agreement (SLA) is not achieved (not achieved) based on the status of the product type (casetype).

From the results of processing and testing using J48 on the dataset, the information is compiled in the form of a tree as shown in Figure 6.

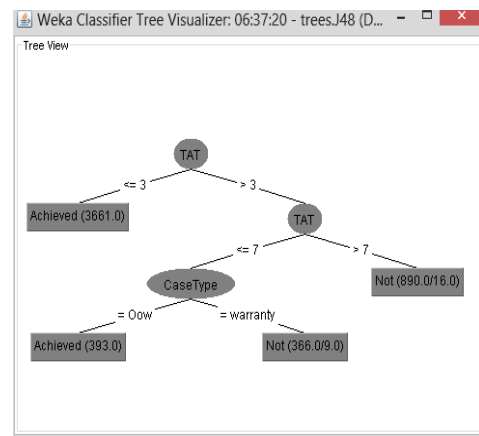


Figure 6. J48 classification model in tree form

From the figure, it can be seen that Turn Around Time (TAT) is the root of the tree. If the Turn Around Time (TAT) can be completed in less than three days (≤ 3) warranty status, the classification results show achieved which means the product unit is repaired in accordance with the agreement on the length of repair with the customer. If the Turn Around Time (TAT) is more than three days (> 3) warranty status, the product unit

is repaired beyond the agreed length of repair, which means that the Service Level Agreement (SLA) was not achieved (not achieved). This does not apply if the Turn Around Time (TAT) is more than three days but less than equal to seven days (≤ 7) out of warranty (oow) status, then the repaired product unit is achieved. And if the product unit that is repaired, Turn Around Time (TAT) is more than seven days (> 7) the status is out of warranty, then the agreement on the service level is not achieved (not achieved).

2) Evaluation using K-Fold Cross Validation

The results of the evaluation resulted in data that were classified correctly (Correctly Classified Instances) in accordance with the grouping of Service Level Agreement (SLA) achieved (achieved) and Service Level Agreement (SLA) not achieved (not achieved) by the algorithm as much as 99.5292% or as many as 5285 data and data that is classified but does not match the predicted class (Incorrectly Classified Instances) which should be in the Service Level Agreement (SLA) group was not achieved (not achieved) but was included in the Service Level Agreement (SLA) group achieved (achieved), which was 0, 4708% or as much as 25 data, as shown in Figure 7. Next is the calculation of the accuracy details.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      5285      99.5292 %
Incorrectly Classified Instances    25        0.4708 %

```

Figure 7. Evaluation results using k-cross validation

True Positive (TP) Rate is data that has been classified as class "x" among all data that actually belongs to class "x". This is equivalent to remembering. In the Confusion Matrix, the diagonal elements are divided by the number of relevant lines, namely $4054 / (4054+25) = 0.99$ SLA class achieved (achieved) and $1231 / (0+1231) = 1$ for SLA class not achieved (not achieved). False Positive (FP) is the proportion of data that has been classified as class "x", but belongs to a different class than previously predicted. In the confusion matrix, the number of column "x" classes under the diagonal element is divided by the number of rows from all other classes, namely $0 / 1231 = 0$ for class SLA achieved (achieved) and $25 / 4079 = 0.006$ class SLA is not achieved (not achieved).

Precision [14] or accuracy is ability to not display inappropriate documents with the needs of the user. In the matrix, the diagonal elements are divided by the number of relevant columns, namely $4054 / (4054 + 0) = 1$ for the SLA class achieved (achieved) and $1231 / (1231+25) = 0, 980$ for the SLA class not achieved (not achieved) . The calculation results from the accuracy details can be seen in Figure 8.

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,994	0,000	1,000	0,994	0,997	0,987	0,996	0,999	Achieved
	1,000	0,006	0,980	1,000	0,990	0,987	0,996	0,971	Not
Weighted Avg.	0,995	0,001	0,995	0,995	0,995	0,987	0,996	0,992	

Figure 8. Class accuracy details

3.4 Analysis of results

The next process will calculate the average percentage of accuracy and error rate in the confusion matrix. The accuracy of the model is calculated using the confusion matrix [15].

Letters a and b in the table respectively indicate the SLA class was achieved (achieved) and the SLA was not achieved (not achieved).

This processing uses 5,310 records of data. Based on the results contained in the confusion matrix, the column represents a prediction, and the row represents the actual class. It can be seen that the number of correct predictions is (aa, bb) and the total number of incorrect predictions is (ab, ba). From the confusion matrix, it can be seen that 4,054 records in class aa are predicted to be correct as class a and as many as 25 records are predicted to be incorrect as the data group class SLA is achieved (achieved), because the record is predicted as class SLA is no bt achieved (not achieved). Furthermore, all records in class bb as many as 1,231 are predicted to be right as class SLA is not achieved (not achieved), from this result can be calculated the average percentage of success accuracy and error rate in the confusion matrix [16, 17].

$$\begin{aligned}
 \text{Accuracy} &= \frac{\text{the number of correct predictions}}{\text{total number of predictions}} \\
 &= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}
 \end{aligned}$$

$$\text{Accuracy} = \frac{4054+1231}{4054+25+0+1231} \times 100\% = 99,5\%$$

$$\begin{aligned}
 \text{Error rate} &= \frac{\text{many wrong predictions}}{\text{total number of predictions}} \\
 &= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}
 \end{aligned}$$

$$\begin{aligned}
 \text{Error rate} &= \frac{25 + 0}{4054 + 25 + 0 + 1231} \times 100\% \\
 &= 0,5\%
 \end{aligned}$$

It can be concluded that the calculation of the percentage level of accuracy in the confusion matrix reaches a percentage value of 99.5% with an error rate or error rate of 0.5%, so that the dataset resolved and canceled cases are declared accurate using the J48 classifier method in determining time accuracy.

4. CONCLUSION

From the results of the classification carried out on the dataset resolved and canceled cases of service center services from January to June 2013, it can be concluded as follows: 1). Determination of the accuracy of repair time by applying the J48 decision tree model which is characterized by the results of data analysis resolved and canceled cases using data mining which is very effective as decision support in services, this can be seen from the percentage of accuracy that reaches more than 99% and percentage error rate / error rate which is only 0.5%. 2. The Turn Around Time (TAT) attribute is the root of the decision tree resulting from the training data and the first determining parameter of the agreement on the length of repair or Service Level Agreement (SLA) for products whose status is still under warranty or is out of warranty period.

REFERENCES

[1] Gudivada, V.N. (2017). Data analytics: Fundamentals. Data Analytics for Intelligent Transportation Systems,

- 31-67. <https://doi.org/10.1016/B978-0-12-809715-1.00002-X>
- [2] Santosa, B., Umam, A. (2018). *Data mining and big data analytics: Theory and Implementation Using Python & Apache Spark*. Penebar Media Pustaka: Yogyakarta.
- [3] De Ville, B., Neville, P. (2013). *Decision trees for analytics: Using SAS Enterprise miner reviews*. Cary, NC: SAS Institute.
- [4] Lesmana, I., Dody, P. (2012). *J48 Decision Tree Development for Diabetes Mellitus Diagnosis*. *Semantik*, pp. 189-193.
- [5] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., Data, M. (2005). *Practical machine learning tools and techniques*. *Data Mining*, 2(4): vii-xiv. <https://doc1.bibliothek.li/acb/FLMF040119.pdf>
- [6] Rokach, L., Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing, Rosewood Drive, Danvers, MA, USA.
- [7] Anon, WEKA, *Data Mining with Machine Learning Group at the University of Waikato*. Available at: <http://www.cs.waikato.ac.nz/ml/weka>.
- [8] Waldburger, M., Stiller, B. (Eds.). (2001). *Definition of a Draft Extended IP Network Management Model*. *IEEE Communications Magazine*, 39(5): 1-36. https://www.simpleweb.org/wiki/images/7/71/Emanics_D8_3.pdf.
- [9] Sherlyanita, A.K. (2017). *Pembuatan Service Level Agreement (SLA) pada Layanan Teknologi Informasi Berdasarkan Kerangka Kerja ITIL V3 2011 (Studi Kasus: DPTSI ITS)* (Doctoral dissertation, Institut Teknologi Sepuluh Nopember). <http://repository.its.ac.id/id/eprint/2468>
- [10] Kusrini, E.T.L., Taufiq, E. (2009). *Algoritma data mining*. Yogyakarta: Andi Offset. <http://diglib.amikom.ac.id/upload/.pdf>.
- [11] Shouman, M., Turner, T., Stocker, R. (2011). *Using decision tree for diagnosing heart disease patients*. In *Proceedings of the Ninth Australasian Data Mining Conference*, 121: 23-30. <https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV121Shouman.pdf>.
- [12] Hulu, S., Sihombing, P. (2020). *Analysis of performance cross validation method and K-Nearest Neighbor in classification data*. *International Journal of Research and Review*, 7(4): 69-73.
- [13] Chou, T.Y., Lei, T.C., Chen, H.H. (2006). *Application of boosting to improve image classification accuracy in rice parcel with decision tree*. In *Proc. 27th Asian Conference on Remote Sensing*, Ulaanbaatar, Mongolia.
- [14] Warnia, N. (2020). *Analisis recall dan precision menggunakan VSM pada kasus text mining*. *InfoTekJar: Jurnal Nasional Informasi dan Teknologi Jaringan*, 5(1).
- [15] Normawati, D., Prayogi, S.A. (2021). *Implementation of naïve bayes classifier and confusion matrix in text-based sentiment analysis on twitter*. *Journal of Computer Science & Informatics (J-SAKTI)*, 5(2): 697-711.
- [16] Zanasi, A., Ruini, F. (2018). *IT-induced cognitive biases in intelligence analysis: big data analytics and serious games*. *International Journal of Safety and Security Engineering*, 8(3): 438-450. <http://dx.doi.org/10.2495/SAFE-V8-N3-438-450>
- [17] Rasheed, M.M., Faieq, A.K., Hashim, A.A. (2020). *Android botnet detection using machine learning*. *Ingénierie des Systèmes d'Information*, 25(1): 127-130. <https://doi.org/10.18280/isi.250117>