JOSCEX Journal of Soft Computing Exploration

Increasing Accuracy of C4.5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease

Aprilia Lestari[1], Alamsyah[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 2, 2020
Revised Aug 20 , 2020
Accepted Sept 12, 2020

Keywords:

Data Mining;
C4.5;
Gain Ratio Boosting;
CKD;

ABSTRACT

Data information that has been available is very much and will require a very long time to process large amounts of information data. Therefore, data mining is used to process large amounts of data. Data mining methods can be used to classify patient diseases, one of them is chronic kidney disease. This research used the classification tree method classification with the C4.5 algorithm. In the pre-processing process, a feature selection was applied to reduce attributes that did not increase the results of classification accuracy. The feature selection used the gain ratio. The Ensemble method used adaboost, which well known as boosting. The datasets used by Chronic Kidney Dataset (CKD) were obtained from the UCI repository of learning machine. The purpose of this research was applying the information gain ratio and adaboost ensemble to the chronic kidney disease dataset using the C4.5 algorithm and finding out the results of the accuracy of the C4.5 algorithm based on information gain ratio and adaboost ensemble. The results obtained for the default iteration in adaboost which was 50 iterations. The accuracy of C4.5 stand-alone was obtained 96.66%. The accuracy for C4.5 using information gain ratio was obtained 97.5%, while C4.5 method using information gain ratio and adaboost was obtained 98.33%.

Corresponding Author:

Aprilia Lestari
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: aprilialestari11@gmail.com

1. INTRODUCTION

Along with the rapid development of technology development, people can easily access information anytime and anywhere. Data information that has been available is very much and will require a very long time to process large amounts of information data. Therefore, to process large amounts of data, data mining techniques are used. Data mining can be applied in various fields. One of them in the health sector is to predict and classify a disease from a patient's medical record data. Data mining methods can be used to classify patient diseases based on the severity of the disease, one of which is to classify patients with kidney disease and not. Chronic kidney disease or kidney failure is a very serious problem in all corners of the world, where the kidneys are damaged and become a cause of non-maximal kidney function [1].

Data mining is used to determine patterns in data mining knowledge and is useful in solving data problems in large data warehouses [2]. The term Data mining is also referred to as knowledge discovery. Data mining has a variety of types of methods, for that the selection of the right method will depend on the purpose and process. One of the methods in data mining is classification. The classification method has input in the form of a collection of records, where each record is marked with tuple (x.y). X is an attribute and Y is a specific attribute / target showing the class label. Classification has several algorithms including Naïve Bayes and C4.5, each of which has different accuracy [3]. Some techniques that exist in classification, decision trees are classification techniques that are very popular and widely used.

Decision tree is the most powerful approach in scientific discovery and data mining, and a very effective tool in various fields such as data and text extraction, information extraction, machine learning, and pattern recognition [4]. One of the most popular decision tree techniques is C4.5. C4.5 algorithm is one of the algorithms developed by J. Ross Quinlan which is the development of algorithm ID3 (Iterative Dichotomiser 3) [5].

This research was conducted using Chronic Kidney Disease Dataset obtained from the UCI repository of learning machine. The following are some of the researches that are relevant to CKD. In the research [6] compared two algorithms namely C4.5 standalone and C4.5 with Pessimistic prunning applied to the Chronic Kidney Disease dataset. Standalone C4.5 has an accuracy of 95% and C4.5 with pressimistic prunning resulting in an accuracy of 96.5%. Based on research [7] discusses the prediction of chronic kidney sufferers using decision tree and naïve bayes algorithms. The dataset used for this research is the chronic kidney disease dataset. The results of this research are decision trees resulting in an accuracy of 91% and Naïve Bayes resulting in an accuracy of 86%. In the research by [8] which states that the C4.5 algorithm has the highest accuracy when applied to the Chronic Kidney Disease dataset compared to the Expectation Maximization (EM) and Artificial Neural Network (ANN) algorithms. The C4.5 algorithm produces an accuracy of 96.75%, EM 70% and ANN 75%.

For improving the accuracy of the C4.5 algorithm, this research used methods in pre- processing and ensemble methods. In the pre-processing process, a feature selection is applied to reduce attributes that do not increase the results of classification accuracy. The feature selection used the gain ratio. The Ensemble method used adaboost which well know as boosting.

The purpose of this research was applying the Information Gain Ratio and Adabost ensemble to the Chronic Kidney Disease dataset using C4.5 algorithm and finding out the results of the accuracy of the C4.5 algorithm based on Information Gain Ratio and Adaboost ensemble.

2. METHOD

2.1 Feature Selection

Feature selection is the process for selecting a subset of original attributes, so that the feature space optimally decreases according to certain criteria. Feature selection which aims to reduce the number of certain features focusing on relevant data and improve quality therefore Feature selection is able to work better than processes that are driven by the selected features [11].

2.2.1 Information Gain Ratio

Information gain ratio is the ratio of obtaining information gain with intrinsic information. To reduce the bias towards multi value attributes by taking the number and size of branches in a calculation when selecting attributes. This is useful as a consideration for logarithmic probabilities to measure the impact of this type of calculation in a dataset.

2.2 Algoritma C4.5

The C4.5 algorithm was introduced by Quinlan as an improved version of ID3. In ID3, induction of decision trees can only be done on categorical type features (nominal / ordinal), while numerical types (internal / ratio) cannot be used. The improvement that distinguishes C4.5 algorithm from ID3 is that it can handle features with numeric types, pruning decision trees, and deriving rule sets. The C4.5 algorithm also uses gain criteria in determining features that are node breakers in the induced tree [12].

In the C4.5 algorithm, building a decision tree the first thing to do is to choose attributes as roots. Then a branch is created for each value in the root. The next step is to divide the case in branches. Then repeat the process for each branch until all the cases in the branch have the same class [13].

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(A)} \quad (1)$$

For calculating the gain, Equation 2 is used as follows [15].

$$\mathrm{Gain}(S, A) = \mathrm{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \mathrm{Entropy}(S_i) \tag{2}$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attribute Partition
$|S_i|$= The Case Number in i Partition
|S| = The Case Number

Meanwhile, calculating the entropy value can be seen in Equation 3.

$$\mathrm{Entropy}(S) = \sum_{i=1}^{n} - p_i * \log_2 p_i \tag{3}$$

Description:
S = Set of case
n = The partition number of S
p_i = The proportion S_i to S, which $log_2 pi$ can be calculated using Equation 4.

$$\log(X) = \frac{\ln(X)}{\ln(2)} \tag{4}$$

Entropy is used to determine which node will be the next training data solver. A higher entropy value will increase the potential for classification. What needs to be considered is that if entropy for nodes is 0 means that all vector data are on the same class label and that node becomes a leaf containing a decision (class label). What also needs to be considered in the calculation of entropy is if one of the elements w_i is 0 then the entropy is confirmed to be 0 too. If the proportion of all w_i elements is equal, it is certain that entropy is worth [16].

For calculate Split Entropy Equation 5 is used as follows.

$$\mathrm{SplitEntropy}_A(S) = -\sum_{i+1}^{n} \frac{|Si|}{|S|} * \log_2 \frac{|Si|}{|S|} \tag{5}$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attribute Partition
$|S_i|$= The Case Number in i Partition
|S| = The Case Number

2.3 Adaptive Boosting (Adaboost)

Adaboost is a boosting algorithm part of ensemble learning that is used to improve classification performance [17]. According to research conducted by Nurzahputra & Muslim [18] states that adaboost is a part of machine learning introduced by Freud and Schapire (1995) which is used to improve accurate prediction rules by uniting many inaccurate and weak regulations.

Adaboost and its variants have been successfully applied in several fields because of their strong theoretical basis, accurate predictions and great simplicity. The steps in the adaboost algorithm are as follows.

a. Input: A collection of research samples with labels {(xi,yi), …, (xn,yn)}, a component learn algorithm, the amount of rotation T.
b. Initialize: Weight of a training sample $w_i^1 = \frac{1}{N}$, for all i=1, …, N
c. Do for t= 1, …, T
    1) Use component learning algorithms to train a classification component, h_t, to weight of a training sample.
    2) Calculate the training error on $h_t = \varepsilon_t \sum_i^N = 1 w_i^t , y_i \neq h_t(x_i)$
        1) Determine the weight for component classifier $h_t = \propto_t = \frac{1}{2} \ln\left(\frac{1- \varepsilon_t}{\varepsilon_t}\right)$

3) Update weight of a training sample $w_i^{t+1} = \frac{w_i^t \exp\{-a_t y_i h_t(x_i)\}}{c_t}, i = 1, \dots, N\ C_t$ is a normalization constant.

4) Output: $f(x) = sign(\sum_t^T = 1\ \propto_t h_t(x))$

3. RESULT AND DISCUSSION

In this research a web-based system was made using the Python programming language to find out the results of applying information gain ratio and Adaboost Ensemble to the C4.5 algorithm in the diagnosis of chronic kidney disease. To make it necessary data related to the diagnosis of chronic kidney disease that will be used as a testing system. This research used a chronic kidney disease dataset obtained from the UCI machine learning repository. This data consists of 24 attributes and 1 class.

At the data processing stage data processing was carried out before the algorithm is applied or commonly called pre-processing. The Chronic kidney disease dataset obtained is in the form of a file with an extension of .arff, changes to the file extension to .xlsx for data processing.

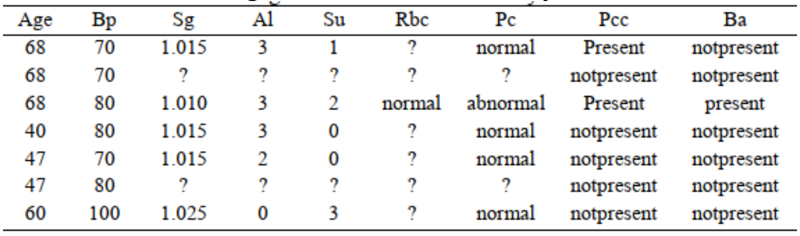
3.1 Formatiing Stage

The following formatting stage is the formatting of standards in the dataset used in the study. For example, in the Rbc (Red Blood Cells) attribute by changing the label on the Rbc attribute to 0 for negative (abnormal) and 1 for positive (normal).

3.2 Handling Missing Value Stage

Handling of missing value is part of pre-processing which aims to optimize mining results. Missing values in datasets are usually marked with the symbol "?" As in Table 1, which is an example of a Chronic Kidney Disease dataset that has a missing value.

Tabel 1. Missing value in the chronic kidney disease dataset



Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba
68	70	1.015	3	1	?	normal	Present	notpresent
68	70	?	?	?	?	?	notpresent	notpresent
68	80	1.010	3	2	normal	abnormal	Present	present
40	80	1.015	3	0	?	normal	notpresent	notpresent
47	70	1.015	2	0	?	normal	notpresent	notpresent
47	80	?	?	?	?	?	notpresent	notpresent
60	100	1.025	0	3	?	normal	notpresent	notpresent

For replacing the missing value in the dataset using the calculation model mean (average) with the Equation 6.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6}$$

a. The value to replace the missing value in attribute Sg:

$$\bar{x}(Sg) = \frac{\sum_1^{359.145} x_{(Sg)}}{353} = \frac{359.145}{353} = 1{,}017$$

b. The value to replace the missing value in attribute Al:

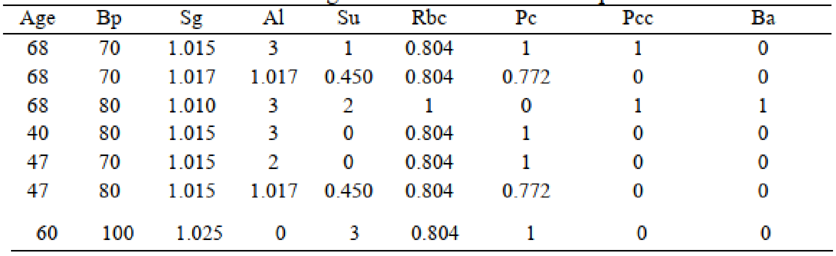
$$\bar{x}(Al) = \frac{\sum_1^{360} x_{(Al)}}{354} = \frac{360}{354} = 1{,}016949$$

c. The value to replace the missing value in attribute Su:

$$\bar{x}(Su) = \frac{\sum_1^{158} x_{(Su)}}{351} = \frac{158}{351} = 0{,}450142 = 0$$

The chronic kidney disease dataset that has been replaced is presented in Table 2.

Tabel 2. The chronic kidney disease dataset



Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba
68	70	1.015	3	1	0.804	1	1	0
68	70	1.017	1.017	0.450	0.804	0.772	0	0
68	80	1.010	3	2	1	0	1	1
40	80	1.015	3	0	0.804	1	0	0
47	70	1.015	2	0	0.804	1	0	0
47	80	1.015	1.017	0.450	0.804	0.772	0	0
60	100	1.025	0	3	0.804	1	0	0

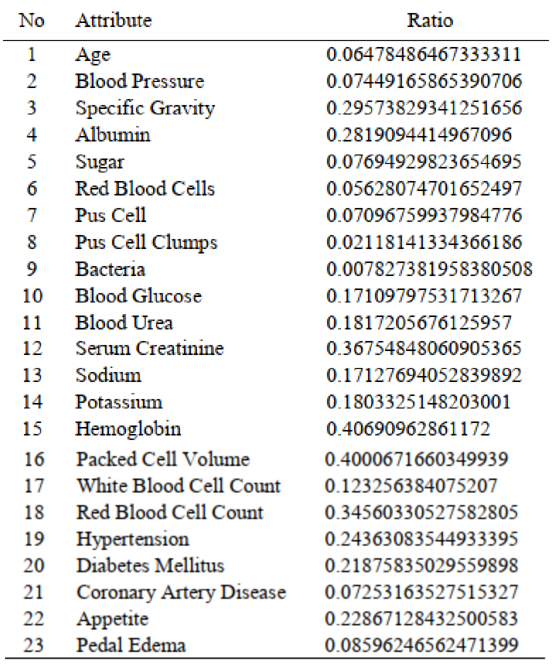
At the stage of class balancing is done by applying the SMOTE algorithm. The SMOTE algorithm is applied to make new data more balanced. German Credit's initial dataset has 1000 samples with 700 loyal (good) classes and 300 churn (bad) classes. Therefore it is necessary to balance the class by creating new data in the churn class. The new dataset of the SMOTE algorithm results in 300 churn class data, so there are 1300 new sample data. This is done so that data can be classified optimally. The attribute selection stage is done by selecting attributes in the data used. In this attribute selection stage there is a dimension reduction in the data in order to optimize attributes that will affect the accuracy of the Naive Bayes algorithm. Dimension reduction in attributes is done by using Genetic Algorithms. Removal of attributes is done one by one from attributes that have the smallest fitness value and will be mining. The process of selecting attributes and mining will stop when the results of the accuracy have exceeded the specified minimum limit.

After going through the pre-processing stage, new data will go through the classification process using the Naive Bayes algorithm. From the results obtained, there is an increase in the accuracy of the Naive Bayes algorithm and the Naive Bayes algorithm by applying the SMOTE algorithm and attibutes selection of Genetic Algorithms.

3.3 Feature Selection Implementation Stage

The stages of applying the feature selection are pre-processing steps in data mining to select features from the original attributes. In this study the application of a feature selection in the chronic kidney disease dataset aims to select attributes that fit certain criteria to improve quality so that optimal results are obtained. The results of information gain ratio for each CKD attribute are shown in Table 3.

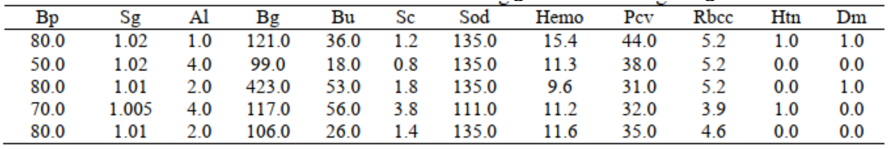
Tabel 3. Result of information gain ratio in CKD



No	Attribute	Ratio
1	Age	0.0647848646733311
2	Blood Pressure	0.07449165865390706
3	Specific Gravity	0.29573829341251656
4	Albumin	0.2819094414967096
5	Sugar	0.07694929823654695
6	Red Blood Cells	0.05628074701652497
7	Pus Cell	0.07096759937984776
8	Pus Cell Clumps	0.02118141334366186
9	Bacteria	0.007827381958380508
10	Blood Glucose	0.17109797531713267
11	Blood Urea	0.1817205676125957
12	Serum Creatinine	0.36754848060905365
13	Sodium	0.17127694052839892
14	Potassium	0.1803325148203001
15	Hemoglobin	0.40690962861172
16	Packed Cell Volume	0.4000671660349939
17	White Blood Cell Count	0.123256384075207
18	Red Blood Cell Count	0.34560330527582805
19	Hypertension	0.24363083544933395
20	Diabetes Mellitus	0.21875835029559898
21	Coronary Artery Disease	0.07253163527515327
22	Appetite	0.22867128432500583
23	Pedal Edema	0.08596246562471399

After the feature selection calculation stage is carried out using the information gain ratio method, the results of the selected attributes are obtained. The results of the feature selection using the information gain ratio method are shown in Table 4.

Table 4. The results of the feature selection using the information gain ratio method



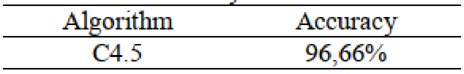
Bp	Sg	Al	Bg	Bu	Sc	Sod	Hemo	Pcv	Rbcc	Htn	Dm
80.0	1.02	1.0	121.0	36.0	1.2	135.0	15.4	44.0	5.2	1.0	1.0
50.0	1.02	4.0	99.0	18.0	0.8	135.0	11.3	38.0	5.2	0.0	0.0
80.0	1.01	2.0	423.0	53.0	1.8	135.0	9.6	31.0	5.2	0.0	1.0
70.0	1.005	4.0	117.0	56.0	3.8	111.0	11.2	32.0	3.9	1.0	0.0
80.0	1.01	2.0	106.0	26.0	1.4	135.0	11.6	35.0	4.6	0.0	0.0

3.3 Data Mining Stage

3.3.1 Implementation of C4.5 Algorithm

In this stage the model used is to apply the C4.5 algorithm to the CKD. New data that is ready to be processed is carried out by sharing training data as a model and testing data to measure the ability of the model formed. In this study the data distribution used a random sub sampling method. Where training data: data testing = 70%: 30% divided randomly. The application of the stand-alone C4.5 algorithm obtained an accuracy of 96.66% is presented in Table 5.

Table 5. The accuracy of C4.5 stand-alone

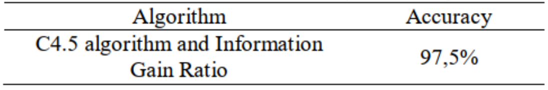


Algorithm	Accuracy
C4.5	96,66%

3.3.1 Implementation of C4.5 Algorithm and Information Gain Ratio

In this stage, the original attribute of the chronic kidney disease dataset consisted of 24 attributes and 1 class, after the information gain ratio method was applied as selecting attributes 12 attributes were selected. In this study the data distribution using the splitter method contained in the sklearn library, the method is random sub sampling. The data sharing system is by sub-sampling random method, where the data is divided into 70% and 30% and the data is taken randomly at each execution. The application of Information Gain Ratio to preprocessing data resulted in an accuracy of C4.5 is 97.5%, the results are presented in Table 6.

Table 6. The accuracy of C4.5 and information gain ratio

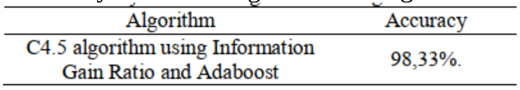


Algorithm	Accuracy
C4.5 algorithm and Information Gain Ratio	97,5%

3.3.1 Implementation of C4.5 Algorithm and Information Gain Ratio and Adaboost

The results of the decision tree will be known the gain value of the attributes that make up the dataset. From the gain value, each attribute is initialized as the initial weight in the calculation of adaboost. After the initialization weight is known, then iterations are determined in adaboost. The default iteration in adaboost is 50 iterations. The accuracy of the C4.5 algorithm based on information gain ratio and adaboost by using sub-random sampling as the splitter is 98.33%. The results of the accuracy obtained are presented in Table 7.

Table 7. The accuracy of C4.5 using information gain ratio and adaboost



Algorithm	Accuracy
C4.5 algorithm using Information Gain Ratio and Adaboost	98,33%.

The results of this accuracy are far better than just using the C4.5 or C4.5 algorithm based on information gain ratio only.

4. CONCLUSION

The application of information gain ratio and adaboost ensemble is a combination of two methods that are useful for increasing accuracy in the C4.5 algorithm. The original attribute of the chronic kidney disease dataset consisted of 24 attributes and 1 class, after the information gain ratio method was applied as selecting attributes 12 attributes were selected. The default iteration in adaboost is 50 iterations. The accuracy of stand-alone C4.5 was 96.66%, for C4.5 with information gain ratio of 97.5%, while C4.5 method was based on information gain ratio and adaboost was 98.33%. So, it can be concluded that combining information gain ratio and adaboost methods can improve classification accuracy.

REFERENCES

[1] Boukenze, B., Mousannif, H., & Haqiq, A. (2016). Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease. International Journal of Database Management Systems (IJDMS), 8(3).

[2] Shajahaan, S. S., Shanthi, S., & ManoChitra, V. (2013). Application Data mining Techniques to Model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering, 3(11): 362-369.

[3] Pranatha, A. A. (2012). Analisis Perbandingan Lima Metode Klasifikasi pada Dataset Sensus Penduduk. Jurnal Sistem Informasi, 4(2): 127-134.

[4] Neeraj, B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data mining. International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), 3(6): 1114-1119.

[5] Muzakir, A., & Wulandari, R. A. (2016). Model Data mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. Scientific Journal of Informatics, 3(1): 19-26.

[6] Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018). Optimization of C4.5 Algorithm-Based Particle Swarm Optimization for Breast Cancer Diagnosis. International Conference on Mathematics, Science and Education, 983(1): 012-063.

[7] Padmanaban, K. A & Parthiban, G. (2016). Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. Indian Journal of Science and Technology, 4(2): 1-5.

[8] S, T., Bai, M., & Majumdar, J. (2017). Analysis and Prediction of Chronic Kidney Disease Using Data Mining Techniques. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 4(9): 25-32.

[9] Gola, J., Britz, D., Staudt, T., Winter, M., Schneider, A. S., Ludovici, M., & Mucklich, F. (2018). Advanced microstructure classification by Data mining methods. Computational Materials Science, 148: 324-335.

[10] Nurzahputra, A., Safitri, A. R., & Muslim, M. A. (2017). Klasifikasi Pelanggan pada Customer Churn Prediction Menggunakan Decision Tree. Prosiding Seminar Nasional Matematika. Semarang: Universitas Negeri Semarang: 717- 722.

[11] Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M., & Mendes, M. P. (2018). Feature Selection Approaches for Predictive Modelling of Groundwater Nitrate Pollution: An Evaluation of Filters, Embedded and Wrapper Methods. Science of the Total Environment, 624(2018): 661-672.

[12] Prasetyo, E. (2014). Data mining: Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi Offset.

[13] Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta: CV Andi Offset.

[14] Neeraj, B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data mining. International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), 3(6): 1114-1119.

[15] Quinland, J. Ross. (1986). Introduction of Decision Tree. Machine Learning. 1(1): 81-106

[16] Han, J. (2012). Data mining Concepts and Techniques. San francisco: Morgan Kauffman.

[17] Listiana, E., & Muslim, M. A. (2017). Penerapan Adaboost Untuk Klasifikasi Support Vector Machine Guna Mengingkatkan Akurasi Pada Diagnosa Chronic Kidney Disease. Prosiding Seminar Nasional Teknologi dan Informatika, 875- 881.

[18] Nurzahputra, A., & Muslim, M. A. (2017). Peningkatan Akurasi pada Algoritma C4.5 Menggunakan Adaboost untuk Meminimalkan Resiko Kredit. Prosiding Seminar Nasional Teknologi dan Informatika. Kudus: Universitas Muria Kudus: 243-247