JurnalMantik
Volume 4, Number 3, Nov 2020, pp. 2119-2126 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Prediction of the Quality of Prospective Student Graduation for Determining New Student Selection Using the C4.5 Decisson Tree Algorithm

Deuis Nur Astrida[1],Gil Diva'ul Haq[2], Alex Azhar Kusuma[3]

[1]Program Studi S1 Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
[2,3]Program Studi SI Teknologi Informasi , Fakultas Ilmu Komputer, Universitas Amikom Purwokerto

Email: deuis@amikompurwokerto.ac.id[1], divapwt123@gmail.com[2], lexse159@gmail.com[3]

ARTICLE INFO

Article history:
Received: 01/11/2020
Revised: 08/11/2020
Accepted: 28/11/2020

Keywords:
Decision Tree, C4.5, Algorithm

ABSTRACT

Historical data of former student contains rich informations. This research begins with a hyphothesis that there must be a correlation between former student data history and their success. By deeply observe the historical background of student, a pattern can be determined. Once a pattern is developed, a prediction of success can be projected to new and current students. This information is helpful for the new student admission strategy. In order to find the pattern a machine learning algorithm will be implemented. C45 works very well for machine learning to generate the rule based on attributes impact to the label. C45 is impelemnted to learn student pattern in the past in order to determine the pattern to predict the future students. A dataset is taken from the last 3 years vocational student graduation dataset. Success is defined as national exam and length of study. A final exam grade of senior high school is consider input to C45 classifier. According to experiments result an accuracy of report score prediction to determine success of national examination score is achieved by C.45 algorithm on success prediction. The test score attribute is taken from presence shows the highest impact to the student success, while attributeachievement as the lowest impact to the output.

1. Introduction

Admission of new students is an important annual agenda for every school. The recurring event every year is the starting point of the process of finding quality resources, namely prospective students. Every school wants to get students with good quality and quantity according to the quota set by the school. The number of prospective students who register must of course be proportional to the capacity of the facilities and infrastructure provided by the school. To overcome this, schools need to make a selection of prospective students who will be accepted. The selection of prospective students includes report card scores, previous achievements, written test results, attendance levels, and interview results.

The quality of prospective students can be detected early by recognizing the patterns and characteristics of prospective students who have existed in previous years and paying attention to quality during the learning period. The information generated can assist management in making student admission decisions in the next academic year.

In previous studies using the neural network (NN) method has advantages in non-linear prediction, strong in parallel and the ability to detect errors, but has weaknesses in the need for large training data, low over-fitting conferences, and its local optimum character [2]. Decision Tree (DT) can solve neural network problems, namely regarding continuous over fitting choosing the right one for attribute selection, regarding training data with missing attributes, and increasing computational efficiency. In general, the success rate of a decision tree is focused on a relatively balanced dataset but has a weakness in the entropy and gain values when the dataset has a high degree and class imbalance.

According to Bisri, using the C4.5 decision tree algorithm which is used for credit risk status. The C4.5 algorithm provides an average level of accuracy of 72.73%. In general, C4.5 is better than k-NN, Random Tree, Naive Bayes and CART, although the accuracy level is still superior to CART results [5].

Agustin Yoga Handoko, using the C4.5 algorithm and adaboost which are used to classify new students. From the experiments carried out, the accuracy value is the same between the C4.5 Algorithm and the Adaboost-based C4.5 Algorithm, namely 77.33% Precision, 90.28% Accuracy, 45.54% Recall but there is a difference in the AUC value for the C4.5 Algorithm of 0.683 while for The Adaboost based C4.5 algorithm is 0.717. This pattern can help to make admissions decisions for new students who can graduate on time and

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

students who graduate late can be predicted earlier by considering the attributes of school origin, UN scores, psychological scores, SPMB scores, and length of study. The difference made by research to be carried out with previous research is the attributes used.

2. Literature review

2.1 Prediction

Errors can be minimized using predictions where this process estimates systematically about something that will happen in the future based on the past. Answers that are generated through predictions do not have to be certain to happen but try to find the results that will be possible [2].

Prediction is also an attempt to predict the future by examining the past. Consists of estimating the future size of several variables such as sales on the basis of knowledge of the past and present. Prediction is used to predict events in the future [4].

2.2 Data Mining

Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases. Data mining is the analysis of observational datasets to find unexpected relationships and to summarize data in new ways that are both understandable and useful for data users [9].

Data mining is an algorithm in digging hidden information in a data collection (database). Data mining analysis runs on data that tends to grow to get the most feasible conclusions and decisions. Datamining has several other names or names, namely: Knowledge discovery (mining) indatabases (KDD), knowledge extraction, data / pattern analysis, business intelligence, etc. [7]

Knowledge Discovery in Database (KDD) is a process of seeking useful knowledge from data. In general, the stages of the process in data mining start with the source data and end with the information generated from several stages. The KDD process is broadly as follows [7]:

a. Data Selection

Selection (selection) of new data from a set of operational data needs to be done before the information mining phase in KDD begins. The selected data that will be used for the data mining process are stored in a separate file from the operational database.

b. Data cleaning (Cleaning)

Before the data mining process can be carried out, it is necessary to carry out a cleaning process on the data which is the focus of KDD. The cleaning process includes removing duplicate data, checking for inconsistent data, and correcting errors in data, such as typographical errors.

c. Transformation

At the transformation stage the data is converted into a form suitable for screening. Some data mining techniques require special data formats before they can be applied. For example, some standard techniques such as association analysis and clustering can only accept categorical data input. Here is also done selecting the data required by the data mining techniques used.

d. Data mining

Data mining is the process of looking for patterns or interesting information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The choice of the right method or algorithm is very much dependent on the objectives and process of KDD as a whole.

e. Interpretation / Evaluation

The information pattern generated from the data mining process needs to be displayed in a form that is easy to understand for interested parties. This stage is part of the KDD process called interpretation. This stage includes examining whether the patterns or information found contradict previous facts or hypotheses?

2.3 Decision Tree C4.5

The C4.5 algorithm is the development of the ID3 algorithm. The algorithms C4.5 and ID3 were created by j. Rose quinlan in the late 1970's. The C4.5 algorithm creates a decision tree from top to bottom, the top attribute is the root, and the bottom one is called the leaf [8].

The decision tree is the most widely used algorithm for classification problems. A decision tree consists of several nodes, namely the tree's roo, internal nods and leafs. The concept of entropy is used to determine which attributes a tree will split. The higher the entropy of a sample, the impure the sample is [3].

The decision quality results obtained from the decision tree method really depend on how the tree is designed. So that if the decision tree is not optimal, it will affect the quality of the decisions obtained [6].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The application of feature selection in this study is to calculate the information gain for each attribute. Information gain of an attribute is obtained from the entropy value before separation minus the entropy after separation [6].

Creating a decision tree is selecting the attributes that should be tested on each node. This process is called information gain which is useful for determining which attributes will be used at each node. The information gain itself is obtained from calculations using a unit called entropy [6].

$$Gain\ (S, A) = Entropy\ Total - \sum_{S}^{Sv} \frac{\text{Nilai Rapor}}{Total} * Entropy\ (Sv) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

3. Research Methodology

The C4.5 algorithm is a well-known algorithm that is used for classification of data that has numeric and categorical attributes. The results of the classification process in the form of rules can be used to predict the value of the discrete type attribute of a new record. Algortima C4.5 itself is a development of the ID3 algorithm, where development is carried out in terms of overcoming missing data, it can handle continuous data and pruning.

In general, the C4.5 algorithm for building a decision tree is as follows:

a. Select attributes as root.
b. Create a branch for each value.
c. Divide cases into branches.
d. Repeat the process for each branch until all cases on the branch have the same class.

To choose the root attribute, it is based on the highest gain value of the existing attributes. To calculate gain, the formula is used as shown in equation 1 below:

Gain (SA) = Entropy (S) - Zi; = i Entropy (Si)

Where :
S: case set
A: attribute
N: the number of partitions attribute A
\ Si \: number of cases on partition i
| Sr | : number of cases in S

The first step the researcher took was to collect data on graduates for the last three years at the school concerned. After that, the data are grouped based on the attributes that have been determined. After the data is grouped correctly, then look for the gain value, according to the following modeling:

The research method used is the action research method. According to Hasibuan (2007) Action research is research that focuses directly on social action. Empowering researchers who go directly to the research area because they cannot be surveyed. The data analysis method used in this research is using case-based reasoning (Case-Based Reasoning). In this reasoning, the knowledge base will contain previously achieved solutions, then a solution will be derived for the current situation (existing facts). This form is used when the user wants to know more in almost the same (similar) cases. In addition, this form is also used when we already have a number of situations or certain cases in the knowledge base. The case-based reasoning method in this study is used to analyze data on new student admissions, teacher ratios and student capacity which can later be used as a basis for researchers in developing the system. The following is the model proposed in Figure 1.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

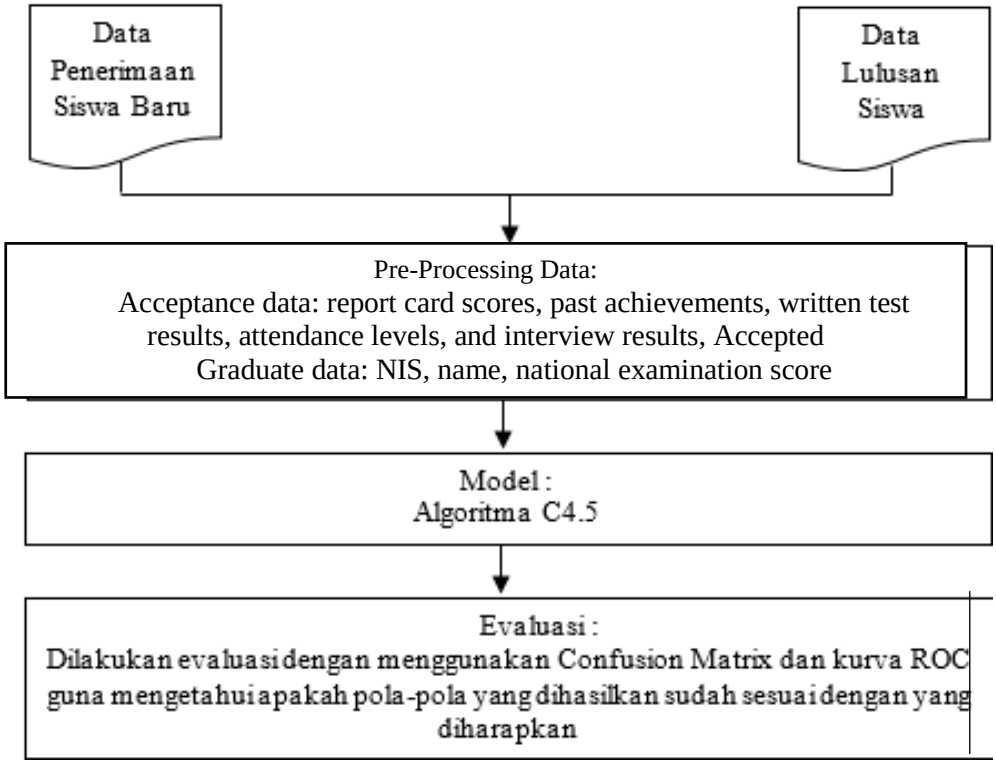

Fig 1. The proposed modeling

The first step carried out by the C4.5 algorithm in Figure 1 is to calculate the gain information for each attribute to determine the top node. Then this step is repeated until all the attributes are found in the nodes in the classification tree. Then the weight with the highest value is selected to obtain a classification tree with a high degree of accuracy. The results of the classification using the C4.5 method will be evaluated using confusion matrix and ROC curves. From the results of the classification, it will produce information that can support decision making for student admissions which will be adjusted to the ratio of teachers and the capacity of students from the school.

4. Results and Discussion

In this study the authors used the Cross-Industry Standard for Data Mining (CRISP-DM) model (P.Chapman, 2000) which consisted of 6 stages, namely: (1) The business understanding stage where the preliminary research was carried out by observing schools to find out directly the conditions and problems that occur. It was found that the results of the UN scores and school rankings based on the UN scores have not been able to reach the predetermined target, this is because it is still difficult to determine the classification of student admission patterns with good accuracy so that a new classification model is needed. (2) The understanding data stage is the stage for obtaining data, where data is obtained from schools from 2017-2020. The data has the attributes of report card scores, achievements that have been achieved, written test results, attendance levels, and the results of interviews, information (accepted / not accepted) as well as NIS graduation data, names, UN scores). From several attributes, the values of the attributes will be grouped in order to get a good classification (table 1).

Table 1.
Attribute Category

Attribute	Score	Category
Report scores for Indonesian, Mathematics and Science (Semester 7-11)	Finish all subject	Very nice
	1 incomplete maple	Enough
	> 2 incomplete maple	Less
Achievement	Ever Champion	Very nice
	There is no	Less
Written test results	Value> = 70	Very nice
	Value <= 70	Less
Attendance rate (Semester 7-11)	Never alpha	Very nice
	Alpha 3 times	Enough
	Alpha> 3 times	Less

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Attribute	Score	Category
Interview	Value> = 70	Very nice
	Value <= 70	Less

(3) The data preparation stage, in this stage the writer takes several steps to prepare the data before processing. The stages carried out are described in Figure 2.

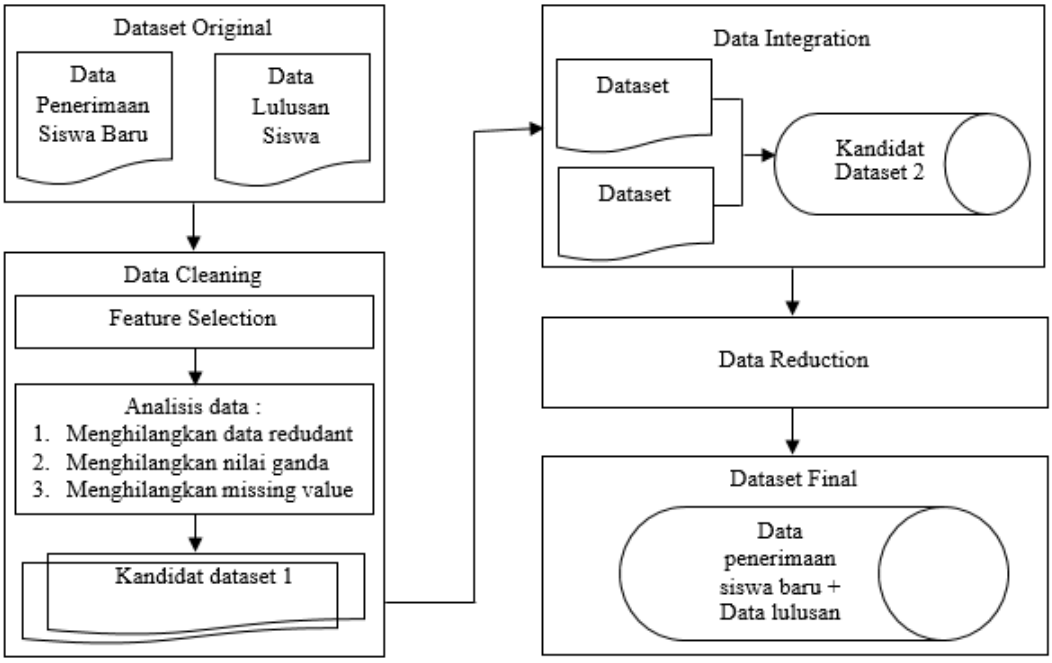

Fig 2. Preparation data

Figure 2 is the process of preparation data starting from taking the original dataset from historical data for new student admissions and historical data on student graduation. This data will be processed in order to obtain the number of data n that has gone through the following stages: (a) Data cleaning, (b) Data integration (c) Data reduction. After this stage is completed, the next stage is (4) Modeling stage (5) Analysis and Pattern Evaluation (6) Deployment phase.

The first step is to perform calculations to find the entropy value and information gain to determine the node to be split. Count the number of student cases accepted and rejected from all cases divided by the attributes of registration no, report card scores, achievements, written test results, attendance, interviews, information received and UN scores. Then the highest Gain value for each Entropy is calculated. The calculation is calculated using the following equation:

= 0.657704779

After that, the entropy calculation of each attribute category is used to obtain the Gain value. Example of calculation for the report card grade attribute with the following equation:

Values (Value Report Card) = Good, Enough, Less

S = [83+, 17-], [S] = 100

S_{Good} = 60

S_{Enough} = 18

S_{Less} = 22

Ent

= 0.122291597

Entro

= 0.309543429

Entr

= 0.845350937

The next stage of the entropy atribu value of the report card will be calculated to find the Gain value using the following equation:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$Gain\ (S, Rapor) = Entropy\ Total - \sum_{i=1}^{n} \frac{\text{Nilai Rapor}}{Total} * Entropy\ (Nilai\ Rapor)$$

$$Gain\ (S, Rapor) = Entropy\ Total - \frac{60}{100} * (S, Bagus) - \frac{18}{100} * (S, Cukup) - \frac{22}{100} * (S, Kurang)$$

$$= 0{,}657704779 - \left(\left(\frac{60}{100}\right) * 0{,}122291597\right) - \left(\left(\frac{18}{100}\right) * 0{,}309543429\right) - \left(\left(\frac{22}{100}\right) 0{,}845350937\right)$$

$$= 0{,}342634797$$

Then with the above equation, do it on all attributes to get information gain to determine the first node to the last node. The results are as follows:

Gain (S, Achievement) = -0,021454873
Gain (S, Test Value) = 0.354990887
Gain (S, Number of Occurrences) = 0.026076818
Gain (S, Interview) = 0.076029544

From the above calculations, it produces a pattern like Figure 3.

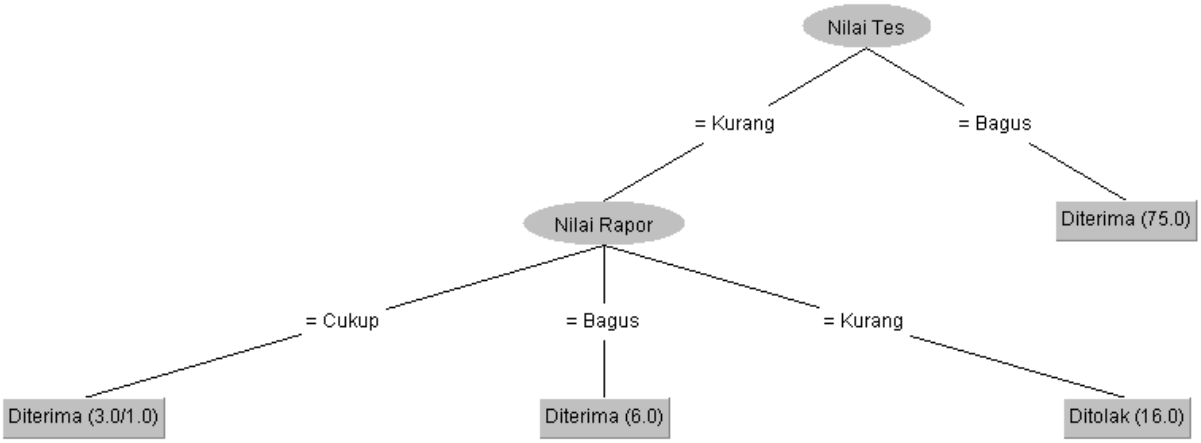

Fig 3. Decision tree model with the C4.5 algorithm

In the decision tree, only the attributes of the test scores and report cards are displayed. This happens because these two attributes are the attributes that have the greatest influence on the acceptance of new prospective students.

The results of the classification algorithm C4.5 using 100 data sets divided into 5 attributes, namely report cards, achievement, test scores, attendance, and interviews. The dataset used shows 83 data with class ACCEPTED and 17 data with class DENIED.

After that, a comparison was made between the student data received with the results of the UN scores that were obtained based on the historical background of the students' UN scores. Based on these comparisons, the conclusion is that students who enter with good report cards and good grades will get good UN scores too.

4.1 Evaluation of the Confusion Matrix Model

True ACCEPTED is a positive tuple in the data set which is classified as positive, amounting to 83 while true DENIED is a negative tuple in the data set which is classified as negative, amounting to 17. False ACCEPTED is a positive tuple in the data set which is classified as negative with 30, while false DENIED is a negative tuple in the data set classified positive amounted to 4. Then from the data above, some equations can be calculated as follows:

Table 2.
Value of Precision, Accuracy and Recall for the C4.5 algorithm

	Score %
Precision	100%
Accuracy	100%
Recall	100%

Table 3.
Confusion Matrix C4.5

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

C4.5	True Received	True Denied
Prediction Accepted	83	17
Prediction Rejected	4	30

From the table above, the level of accuracy can be calculated as follows:

$$\text{Accuracy} = \left(\frac{83+30}{83+17+4+30}\right) * 100\% = 84{,}32\%$$

4.2 ROC curve

The ROC curve shows the trade-off between the degree to which a model can accurately recognize positive data and the degree to which the model mistakes negative data as positive data. To measure the accuracy of a model, we can measure the area under the ROC curve.

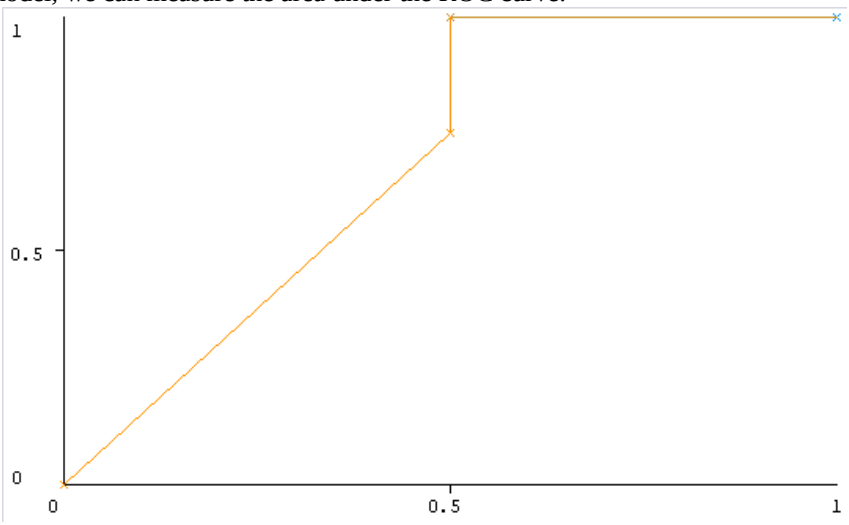

Fig 4.C4.5 Algorithm Accuracy Curve

Figure 4 shows the ROC graph with the AUC (Area Under Curve) value with C 4.5 of 0.8875. The AUC accuracy is said to be perfect if the AUC value reaches 1,000 and the accuracy is bad if the AUC value is below 0.500, the following is the AUC value classification table (table 5).

Table 4.
AUC Value Classification Table

AUC value	Classified as
0.90 - 1.00	Excellent
0.80 - 0.90	Good
0.70 - 0.80	Fair
0.60 - 0.70	Poor
0.50 - 0.60	Fail

5. Conclusion

Based on the research conducted, the following conclusions can be drawn:

a. The C4.5 algorithm is successful in predicting prospective students with an accuracy rate of 84.32%.
b. The most influential factor is test scores while the least influential factor is achievement.
c. The more factors used for the C4.5 algorithm, the greater the accuracy value.

Thank-you note

Our gratitude goes to Amikom Purwokerto University and LPPM Amikom Purwokerto University with the assistance of Amikom Mitra Masyarakat grant. Our gratitude also goes to the Principal of Wangon Public High School and the teachers and students involved.

6. Reference

[1] Agustin, Y. H., . K., & Luthfi, E. T. (2017). Klasifikasi Penerimaan Mahasiswa Baru Menggunakan Algortima C4.5 Dan Adaboost (Studi Kasus : STMIK XYZ). *CSRID (Computer Science Research and Its Development Journal)*, *9*(1), 1. https://doi.org/10.22303/csrid.9.1.2017.1-11

[2] Ali, I., & Sularto, L. (2019). Optimasi Parameter Artificial Neural Network Menggunakan Algoritma Genetika Untuk Prediksi Kelulusan Mahasiswa. *Jurnal ICT : Information Communication & Technology*, *18*(1), 54–59. https://doi.org/10.36054/jict-ikmi.v18i1.52

[3] Amelia, M. winny, Lumenta, A. S. ., & Jacobus, A. (2017). Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes. *Jurnal Teknik Informatika*, *11*(1). https://doi.org/10.35793/jti.11.1.2017.17652.

[4] Bakker, Anton Sutan Takdir Alisjahbana, M. A. A., Kontributor:, & Toety Heraty Noerhadi, J. Sudarminta, P. Hardono Hadi M. Mukhtasar Syamsuddin, R. A. . W. (2011). *M etodologi p enelitian f ilsafat*. 16–42.

[5] Bisri, A. (2015) 'Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree', *Journal of Intelligent Systems*, 1(1), pp. 27–32.

[6] Capparuccia, R., De Leone, R., & Marchitto, E. (2007). Integrating support vector machines and neural networks. *Neural Networks*, *20*(5), 590–597. https://doi.org/10.1016/j.neunet.2006.12.003

[7] Hormann, A. M. (1964). Programs for machine learning. Part II. *Information and Control*, *7*(1), 55–77. https://doi.org/10.1016/S0019-9958(64)90259-1

[8] Kumar, B. and Pal, S. (2011) 'Mining Educational Data to Analyze Students Performance', International Journal of Advanced Computer Science and Applications, 2(6). doi: 10.14569/IJACSA.2011.020609.

[9] Sri Kusumadewi. (2003). Rtificial Ntelligence. *Artificial Intelligence (Teknik Dan Aplikasinya*).

[10] Romadhona, A., Suprapedi , S. dan Himawan, H. (2017). Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma Decision Tree. *Jurnal Teknologi Informasi*, *13*, 69–83.