Jurnal CoreIT, Vol.9, No.2, December 2023
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Comparison Of The Performance Of C4.5 And Naive Bayes Algorithms For Student Graduation Prediction

Baskoro1*), Bambang Triraharjo2), Adi Wibowo)
1 *Faculty of Science and Technology, Pringsewu Muhammadiyah University,* 2 *Faculty of Science and Technology, Pringsewu Muhammadiyah University,* 3 *Faculty of Engineering and Computers, Kotabumi Muhammadiyah University*
1st*baskoro@umpri.ac.id*(*Corresponding Author)., 2nd*bambangtriraharjo@umpri.ac.id*., 3rd*adi.wibowo@umko.ac.id*.,

Abstract. *Along with the development of technology, especially the development of increasingly large data storage. One organization that has large data storage is an educational organization. Educational organizations use data to obtain information, especially information about students. Student data has many attributes so that we can make predictions such as predictions of student performance, predictions of scholarship recipients and predictions of student graduation. Data mining methods in education are classified into five dimensions, one of which is prediction, such as predicting output values based on input data. From the results of the research conducted from the initial stage to the testing stage of the application of the C4.5 Algorithm, the accuracy results are higher than Naïve Bayes because in its classification stage, C4.5 processes attribute data one by one. The difference is with naïve Bayes which is influenced by the amount of data used, the comparison of the amount of training and testing data. The feasibility of the model obtained is supported by the high accuracy, precision, recall and AUC obtained from the two algorithms that have been tested. The C4.5 algorithm has an accuracy rate of 79.91%, 89.06% precision and 81.38% recall and an AUC value of 0.823. Meanwhile, Naïve Bayes has an accuracy rate of 76.95%, precision of 75.95% and recall of 98.38% and an AUC value of 0.838.*

Keywords: *Graduation, Prediction, Data Mining, C4.5, Naïve Bayes*

Received July 2023 / **Revised** December 2023 / **Accepted** December 2023

INTRODUCTION

Along with the development of technology, especially the development of increasingly large data storage. Data is a repository of information that can be used to analyze organizational needs[1]. One organization that has large data storage is an educational organization. Educational organizations use data to obtain information, especially information about students. Student data has many attributes so that we can make predictions such as predictions of student performance, predictions of scholarship recipients and predictions of student graduation.

Academic achievement is the main thing that is used as a parameter of educational success. One indicator of achieving these goals is the results of student academic achievement as expressed by the Semester Grade Point Average (IPS) and the Grade Point Average (GPA)[2]. Semester Achievement Index is the value of student academic achievement with all courses taken in each particular semester. And the Cumulative Grade Point Average is a student's academic achievement by combining all courses taken up to a certain semester. Realizing quality education related to the role of lecturers, student motivation, student discipline, student socio-economic as well as past learning outcomes. This data has the potential to yield useful new information. One of the things that can be done by data mining is to predict student academic achievement [3]. If student academic achievement can be known earlier, then the study program can take the necessary actions so that students can achieve good academic achievement. The final hope is that all students from various backgrounds can maximize their academic achievements [4]. Based on the explanation above, the focus of this study is to predict student academic achievement using the data mining classification method based on the role of lecturers, student motivation, student discipline, student socio-economics and also previous learning outcomes.

DOI : 10.24014/coreit.v9i2.24931

Graduation studies and student achievement are very important for both students, parents and study programs. Many studies have been developed regarding the prediction of student achievement, including research conducted by Hendra, Mochammad Abdul Azis and Suhardjono with the title Analysis of Student Graduation Predictions Using Particle Swarm Optimization Based Decision Trees. The attribute used in this study as an indicator of student achievement is semester social studies [5].

Student study research was also conducted by Aryasanti on predicting student study failure using the Naive Bayes and Decision Trees algorithms. The results found that the Naïve Bayes algorithm provides better accuracy than decision trees [6]. Research was also carried out by Budiantara et al, namely with the title Comparison of Decision Tree Algorithms, Naive Bayes and K Nearest Neighbors to Predict Student Graduates on Time. Based on the results of the study, it was found that the accuracy value of the C4.5 and Naïve Bayes algorithms is almost the same, which is above 95% [7].

The importance of predicting student graduation at an institution encourages this research to be carried out. In this study selected using the datamining algorithm. Data mining in the field of education is not like data mining in general because the data hierarchy is different from other fields. Data mining methods in education are classified into five dimensions, one of which is prediction, such as predicting output values based on input data. Predict data mining there are several data mining techniques using algorithms such as naïve bayes, decision trees, K-nearest neighbors, neural networks. naïve bayes because naive bayes is the 10 best ranking algorithm so it can be used in decision making [8]. The Naïve Bayes method is a method that can be used in decision making to get better results on a classification problem and the Naive Bayes algorithm method is used for the performance of the naïve Bayes classification which has a high enough ability to predict future opportunities based on experience or data in the past[9]. Based on several previous studies it was found that Naïve Bayes is a method that provides a better level of accuracy compared to comparison algorithms, so in this study the prediction of student graduation using data mining is by looking at the variables that affect the goodness of the model, namely the first 2 semester variables or the first 4 semesters of students. Furthermore, the steps and variables that influence the determination of student graduation are discussed in the research methodology.

METHODS

A. Research Flow

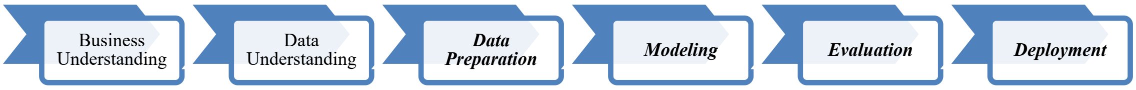


Figure 1 Research Flow

1. Understanding the Business Phase
   At this stage it focuses on research objectives, namely to find out the best algorithm for predicting student graduation by translating student academic historical data from the Bureau of Academic and Student Administration (BAAK) Muhammadiyah Pringsewu University, so that the best model is obtained to fulfill the research objectives.
2. Data Understanding Phase
   The data to be used in the research is data from the results of collecting student academic historical documentation data from the Bureau of Academic and Student Administration (BAAK) Muhammadiyah Pringsewu University.
3. Data Preparation
   To facilitate understanding of the instrument, data preprocessing was carried out.
4. Modeling (Modeling Phase)
   The algorithms used in this study are the C4.5 and Naive Bayes algorithms to classify in predicting student graduation at Pringsewu Muhammadiyah University and to obtain a model or function to describe graduation predictions by comparing the C4.5 and Naive Bayes algorithms.
5. Evaluation Phase
   At this stage, the performance evaluation of the two algorithms is carried out, namely the C4.5 Algorithm and Naive Bayes by comparing the results of the average values of accuracy, recall, and error rate found in the confusion matrix table.
6. Deployment Phase

DOI : 10.24014/coreit.v9i2.24931

After the evaluation stage where the results of a model are assessed in detail, the implementation of the entire model that has been built is carried out.

B. Decision tree (C4.5)

C4.5 is a collection of algorithms for classification techniques in machine learning and data mining. The goal is supervised learning, where each tuple in a data set can be described by a set of attribute values, and each tuple belongs to one of many different and incompatible classes[10]. The objective of C4.5 is to study the mapping from attribute values to categories that can be used to categorize unknown items into new categories. J. Rossi Quinlan suggests C4.5 based on ID3. A decision tree is built using the ID3 algorithm. A decision tree is a flowchart-like tree structure, with each internal node (nonleaf node) representing a test on an attribute, each branch representing a test result, and each leaf node holding a class label. After constructing a decision tree for tuples for which no classification label is provided, we choose a path from the root node to a leaf node, and that path holds the tuple's prediction information. Decision trees have the advantage of not requiring domain information or parameter configuration, making them ideal for exploratory information mining.

The C4.5 algorithm is based on ID3 added to continuous attributes, attribute values, and information processing, by generating a tree to build a pruning decision tree in two stages. On each attribute by calculating the information algorithm C4.5, we can find out the Gain Ratio the rate of information acquisition [11]. Finally, it is selected with the highest information acquisition rate from the attribute test set given to set branch. According to the test attribute values using a recursive algorithm, get the initial decision tree. The computational formula related to the C4.5 algorithm is as follows [12]. First, the expected values required for sample classification are given as follows: Determine the root of the tree by calculating the highest gain value of each attribute or the lowest entropy index value. Previously, the entropy index value was calculated using the formula:

$$Entropy\ (i) = \sum_{j=1}^{m} \quad f(i,j).\ 2f[(i,j)] \tag{1}$$

a. Gain value with the formula:

$$gain = -\sum_{i=1}^{p} \qquad .\ IE(i) \tag{2}$$

b. To calculate the gain ratio, it is necessary to know a new term called Split Information with the formula:

$$SplitInformation = -\sum_{t=1}^{c} \frac{\mathrm{S1}}{\mathrm{S}} log2 \frac{S1}{S} \tag{3}$$

c. Next calculate the gain ratio:

$$Gainratio(S,A) = \frac{\mathrm{Gain(S.A)}}{\mathrm{SplitInformation\ (S,A)}} \tag{4}$$

d. Repeat step 2 until all records have been split. The decision tree splitting process ends when:

1) All tuples in m node record are of the same class.
2) Attributes in the dataset are not subdivided.
3) An empty branch has no records

C. Naïve Bayes

Bayesian classification is a statistical classification that can be used to predict the probability of membership in a class discovered by British scientist Thomas Bayes[13]. Naive Bayes is a classification algorithm that is quite simple and easy to implement so that this algorithm is very effective when tested with the correct data set, especially if Naive Bayes is combined with function selection, so Naive Bayes can reduce redundancy in data, besides that Naive Bayes shows good results when combined with the clustering method [14]. Naive Bayes is proven to have high accuracy compared to support vector machines.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{5}$$

Then X is evidence, H is hypothesis, P(H|X) is probability is hypothesis H is true evidence X or at P(H|X) is posterior probability H with condition X, P(X|H) is probability is evidence X true or hypothesis H or Posterior probability X equal to the condition H ,P(H) is the prior probability of hypothesis H, and P(X) is the prior probability of proof X.

$$P(C|F1 \ldots\ldots Fn) = \frac{P(C)P(F1.Fn|C)}{P(F1 \ldots Fn)} \tag{6}$$

So the C variable describes the class, while the F1...Fn variable describes the guiding character in classifying. Where this formula describes the probability that the sample includes a special character in class C (Posterior), namely the probability of leaving class C (before the entry of the sample, many priors are made), multiplied by the probability that class C sample characters appear (also called the likelihood), divided by the probability of the character appearing global examples (also called evidence) [15]. The above formula can be made simply as follows.

$$Posterior = \frac{Prior\ x\ likelihod}{evidence} \tag{7}$$

The continuous data classification uses the Gauss Density formula:

$$P(Xi = Xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu i)^2}{2\sigma 2ij}} \tag{8}$$

Where :

Q: Opportunity
Xi : Attribute to i
xi : Value of attribute to i
Y : Class searched
yi : Subclass Y searched
μ : mean, describes the average of all attributes
σ :Standard deviation, describes the variance across all attributes.

D. Performance Evaluation

1. Confusion matrix

This method only uses matrix tables as in Table 1, if the dataset consists of only two classes, one class is considered positive and the other is negative. Evaluation with the confusion matrix produces accuracy, precision, and recall [16].

Table 1 Confusion Matrix

Correct	*Classified as*	
Classification	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

True Positive is the number of positive records which are classified as positive, false positive is the number of negative records which are classified as positive, false negative is the number of positive records which are classified as negative, true negative is the number of negative records which are classified as negative.

2. Receiver operating characteristic (ROC)

The ROC curve is a graphical plot illustrating the diagnostic capability of a binary classifier system as its discrimination threshold varies. This method was originally developed for military radar receiver operators starting in 1941, hence its name. The ROC curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate is also known as sensitivity, recall, or probability of detection. The false positive rate is also known as the probability of a false alarm and can be calculated as (1 - specificity)[17]. ROC can also be thought of as a plot of power as a function of the Type I Error of the decision rule (when performance is calculated from only a sample of the population, it can be thought of as an estimator of this sum). The performance accuracy of AUC can be classified into several groups, namely [18]:

1. 0.90 – 1.00 = Excellent Classification
2. 0.80 – 0.90 = Good Classification
3. 0.70 – 0.80 = Fair Classification
4. 0.60 – 0.70 = Poor Classification
5. 0.50 – 0.60 = Failure Classification

RESULT AND DISCUSSION

A. Classification

This process is an implementation of making a classification model in classifying data. In this process there are two stages, namely the formation of a tree and changing the tree into a rule. In this process, the Rapid miners application is used as a tool to create data mining processes. In the Decision Tree algorithm, records that have been imported into the rapid miner are used to determine the decision tree pattern. The application of data to Rapid Miner is used to predict student graduation using the Decision Tree algorithm shown in Figure 2 below:

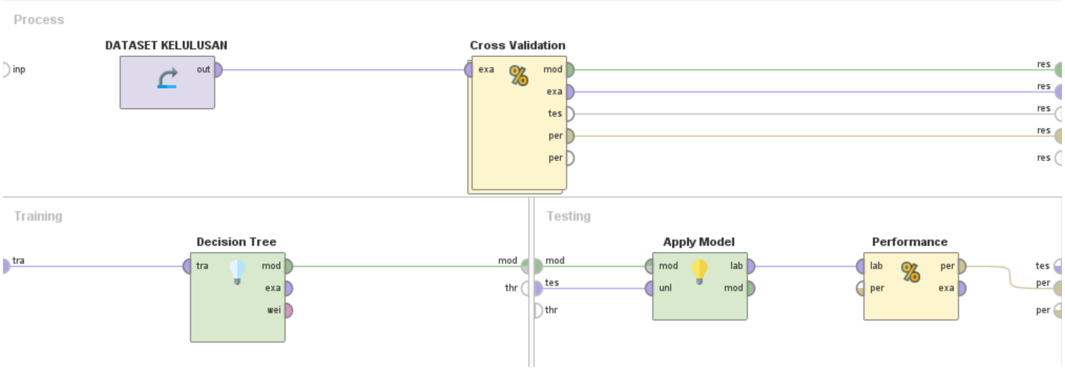
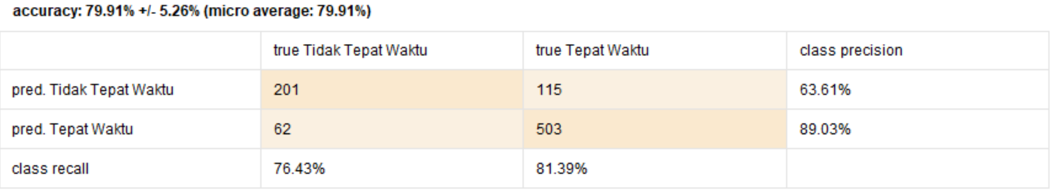


Figure 2 C4.5 Algorithm Classification Model

After carrying out the above steps in the classification process for the C4.5 algorithm method, a model formed from the C4.5 algorithm classification process will be obtained in the form of a decision tree. Decision tree is formed based on nodes. The nodes in the decision tree are the variables used in the research. Data testing using the C4.5 algorithm also obtained an accuracy result table as shown in Figure 4.5 below. We can see the test results in Figure 3 below.


accuracy: 79.91% +/- 5.26% (micro average: 79.91%)

	true Tidak Tepat Waktu	true Tepat Waktu	class precision
pred. Tidak Tepat Waktu	201	115	63.61%
pred. Tepat Waktu	62	503	89.03%
class recall	76.43%	81.39%	

Figure 3 Table of results of the accuracy of testing the C4.5 Algorithm

From Figure 3 it can be concluded that the accuracy level of the C4.5 algorithm method is very high, reaching 79.91%, where the amount of data that is predicted is not on time and in fact is not on time is 201, the amount of data that is predicted is on time and in fact is not on time is as much as 115, the amount of data that is predicted to be not on time and in fact is on time is 62, and the amount of data that is predicted to be on time and in fact is on time is 503.

The classification process using the Naive Bayes model is used to describe or predict opportunities based on each condition. In this process, the rapid miners application is used as a tool to create a data mining process. The following is an illustration of the implementation of the Naive Bayes model using rapid miner.

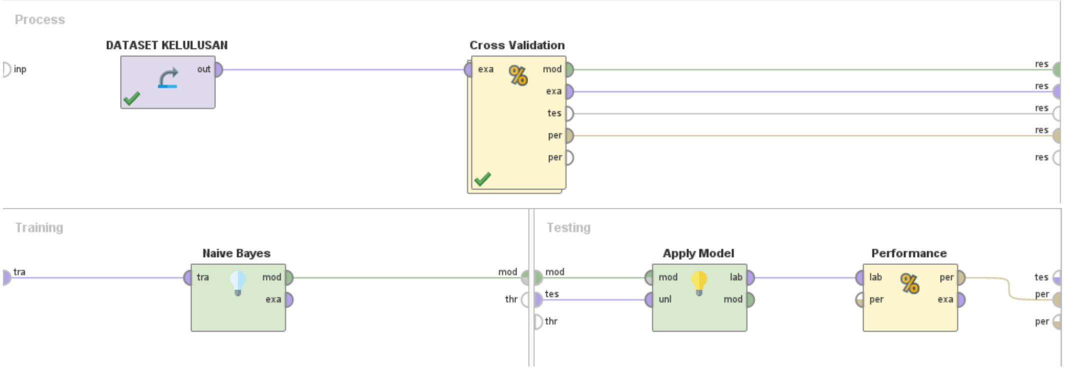
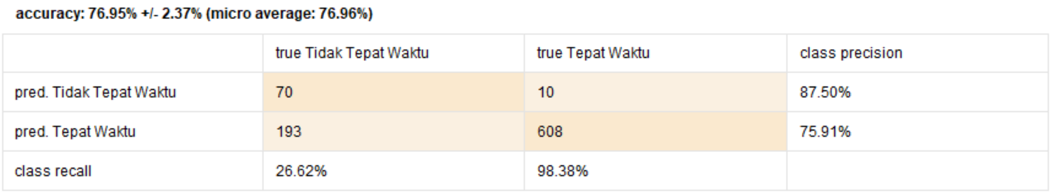


Figure 4 Naive Bayes Algorithm Classification Model

Based on Figure 4 which has been built on the rapid miner application, the following results are obtained:

DOI : 10.24014/coreit.v9i2.24931



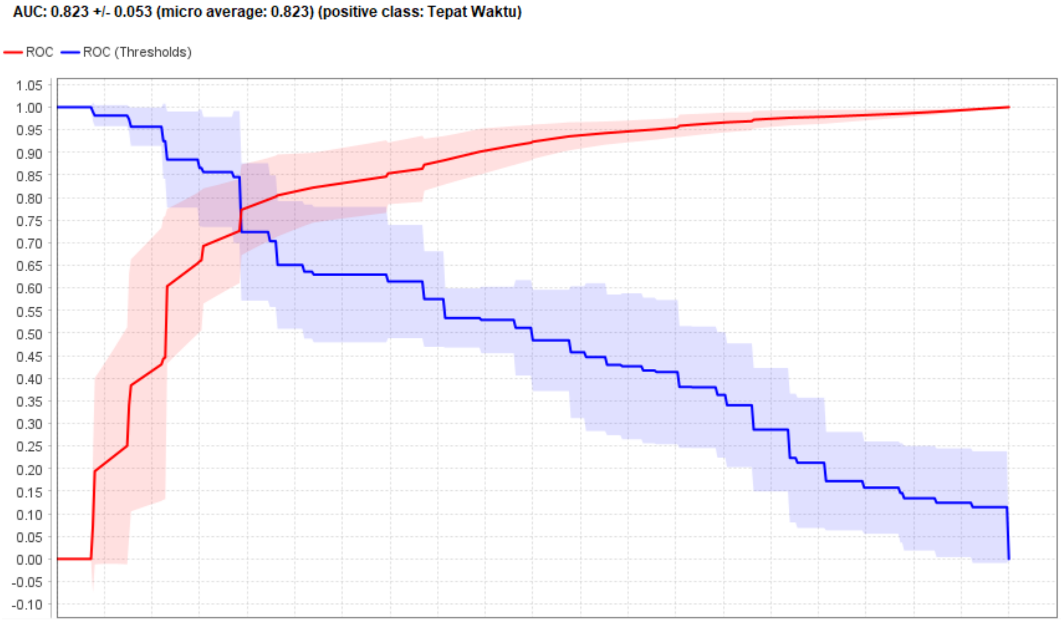
accuracy: 76.95% +/- 2.37% (micro average: 76.96%)

	true Tidak Tepat Waktu	true Tepat Waktu	class precision
pred. Tidak Tepat Waktu	70	10	87.50%
pred. Tepat Waktu	193	608	75.91%
class recall	26.62%	98.38%	

Figure 5 Table of results of the accuracy of testing the Naive Bayes Algorithm

From Figure 5 it can be seen that the data testing carried out using the Naive Bayes model has a fairly high level of accuracy, namely 76.95%, this shows that the classification process is good. where the amount of data that is predicted to be not on time and in fact not on time is 70, the amount of data that is predicted to be on time and in fact is not on time is 10, the amount of data that is predicted is not on time and in fact is on time is 193, and the amount of data that is predicted is right time and in fact there are 608 times.

In addition to the Confusion matrix to determine the performance of this experiment, we can rely on the resulting AUC curve. We can see the comparison of the results of the AUC curve using the C4.5 and Naïve Bayes algorithms in Figures 6 and 7 below:



AUC: 0.823 +/- 0.053 (micro average: 0.823) (positive class: Tepat Waktu)

Figure 6 The AUC curve uses the C4.5 algorithm

The ROC curve in Figure 6 shows the accuracy results and visually compares the classification with false positives as horizontal lines and true positives as vertical lines. From Figure 6 is a visualization of the AUC 0.823 results obtained by Algorithm C4.5. As for the Naïve Bayes ROC results, it can be seen in Figure 7 below

DOI : 10.24014/coreit.v9i2.24931

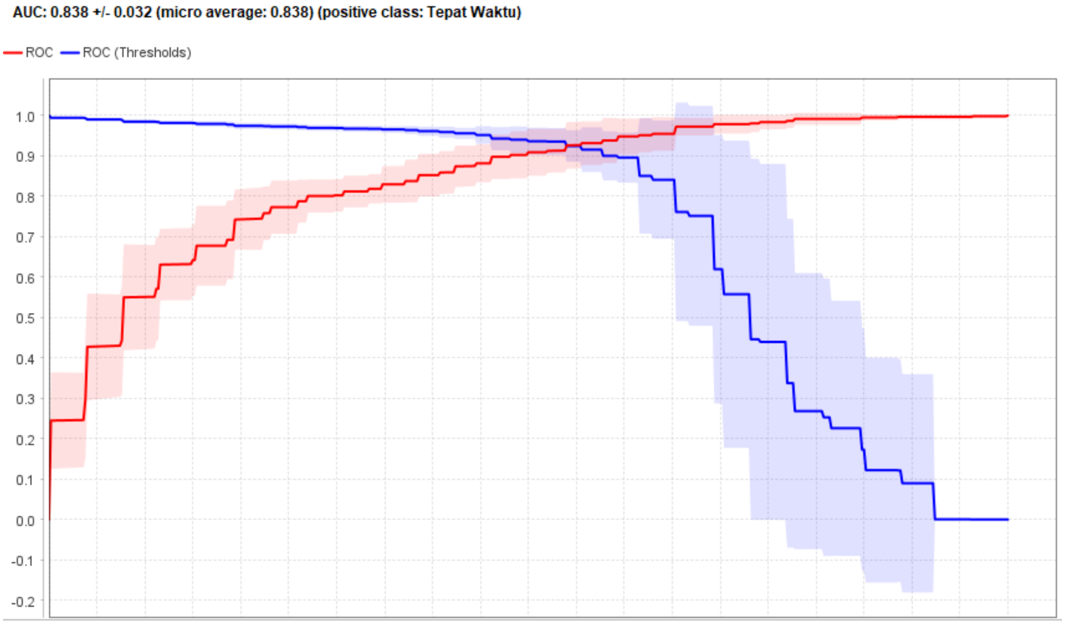


Figure 7 The AUC curve uses the Naïve Bayes algorithm

The Naïve Bayes ROC curve shown in Figure 7 shows a visualization of the AUC 0.838 results which are included in the Good Classification category.

B. Evaluation

Based on the results of the tests that have been carried out, the model is suitable for use as a predictive model for student graduation on time, the feasibility of the model obtained is supported by the level of accuracy of the two models used in this study which are shown in table 2 below.

Table 2 Algorithm performance comparison

Metode	*accuracy*	*precision*	*recall*	AUC
C4.5	79.91%	89.06%	81.38%	0.823
Naïve Bayes	76.95%	75.95%	98.38%	0.838

Based on table 2 it can be seen that the C4.5 Algortma method is superior to Naïve Bayes. The results of the analysis of the C4.5 algorithm model have an accuracy rate of 79.91%, an AUC value of 0.823, a precision level of 89.06% and a recall of 81.38%. Meanwhile, Naïve Bayes has an accuracy rate of 76.95%, an AUC value of 0.838, a precision level of 75.95% and a recall of 98.38%.

CONCLUSION

Based on the discussion that has been described, the C4.5 algorithm gets higher accuracy results than Naïve Bayes because in its classification stage, C4.5 processes attribute data one by one. The difference is with naïve Bayes which is influenced by the amount of data used, the comparison of the amount of training and testing data. The feasibility of the model obtained is supported by the high accuracy, precision, recall and AUC obtained from the two algorithms that have been tested. The C4.5 algorithm has an accuracy rate of 79.91%, 89.06% precision and 81.38% recall and an AUC value of 0.823. Meanwhile, Naïve Bayes has an accuracy rate of 76.95%, precision of 75.95% and recall of 98.38% and an AUC value of 0.838. This method can be used to predict student graduation and assist the university in mapping student graduation. For future researchers, try using applications other than Rapidminner in data analysis and try using other methods besides C4.5 and Naive Bayes. Then add more records and attributes and parameters in data processing and data need to adapt to the latest curriculum. A graph is made of the number of graduates each year to find out whether there is an increase or not.

REFERENCES

[1] Bruce Ratner, “Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data Third Edition,” 2017.

[2] Eko Prasetiyo Rohmawan, “PREDICTION OF STUDENT GRADUATION ON TIME USING DESICION TREE METHOD AND ARTIFICIAL NEURAL NETWORK,” 2018.

[3] S. Novia Hermawanti and A. Adi Sunarto, "IMPLEMENTATION OF C4.5 ALGORITHM FOR PREDICTING ON TIME GRADUATION (Case Study: Informatics Engineering Study Program)," *Jurnal Ilmiah SANTIKA*, vol. 9, no. 1, 2019.

[4] U. Kristen *et al.*, "Manage the Journal of Education Management Master of Education Management FKIP," no. 1, pp. 74–85, 2018.

[5] R. Mikut and M. Reischl, "Data mining tools," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 431–443, Sep. 2011, doi: https://doi.org/10.1002/widm.24.

[6] B. Seref and E. Bostanci, "Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1–7. doi: 10.1109/ISMSIT.2018.8567243.

[7] T. Sinta Peringkat *et al.*, "COMPARISON OF DECISION TREE, NAIVE BAYES AND K-NEAREST NEIGHBOR ALGORITHMS FOR PREDICTING STUDENTS TO GRADUATE ON TIME," 2020, [Online]. Available: www.bri-institute.ac.id

[8] F. D. Pranatasari, "THE INFLUENCE OF ACADEMIC SUPERVISOR MENTORING ON STUDENT ACADEMIC ACHIEVEMENT," 2016. [Online]. Available: http://forlap.dikti.go.id/,

[9] A. Pratama, R. Cahya Wihandika, and D. E. Ratnawati, "Implementation of Support Vector Machine (SVM) Algorithm for Predicting Student Graduation Timeliness," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[10] Parteek Bhatia, "Data Mining and Data Warehousing," 2019.

[11] D. Forsyth, "Probability and Statistics for Computer Science," 2018.

[12] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4.5 algorithm for english emotional classification," *Evolving Systems*, vol. 10, no. 3, pp. 425–451, Sep. 2019, doi: 10.1007/s12530-017-9180-1.

[13] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 403–412. doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

[14] O. Caelen, "A Bayesian Interpretation of the Confusion Matrix," 2017.

[15] M. Kubat, *An Introduction to Machine Learning*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-63913-0.

[16] J. Unpingco, *Python for probability, statistics, and machine learning*. Springer International Publishing, 2016. doi: 10.1007/978-3-319-30717-6.

[17] D. J. H. Wojtek J. Krzanowski, "ROC Curves for Continuous Data," 2009.

[18] J. Moolayil, *Learn Keras for Deep Neural Networks*. Apress, 2019. doi: 10.1007/978-1-4842-4240-7.