JIKO (Jurnal Informatika dan Komputer)
Vol. 7, No. 2, August 2024, pp. 117-122
DOI: 10.33387/jiko.v7i2.8506

Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
p-ISSN: 2614-8897
e-ISSN: 2656-1948

COMPARISON OF DECISION TREE AND NAÏVE BAYES ALGORITHMS IN PREDICTING STUDENT GRADUATION AT YPK IMANUEL JUNIOR HIGH SCHOOL, NABIRE REGENCY

Kristia Yuliawan [1], Stevanus Murib [2]

[1,2] Computer Science Study Program, STMIK Pesat Nabire
*Email: [1]christianpesat@gmail.com, [2]st3v4nusmur1b@gmail.com

(Received: 24 July 2024, Revised: 2 August 2024, Accepted: 16 August 2024)

Abstract

This study aims to compare the accuracy of the Decision Tree C4.5 and Naive Bayes algorithms in predicting student graduation at YPK Immanuel Nabire Junior High School, Central Papua. Student data from the 2022 and 2023 school years were used as training data, whereas student data for the 2024 school year were used as testing data. Data collection methods included field studies, interviews with schools, and literature studies. The implementation of the algorithm is carried out using the Orange software, which simplifies the process of data visualization and analysis. Both algorithms are applied to data processed through stages of cleaning and normalization to ensure the quality and relevance of the data used. The results show that the Decision Tree C4.5 algorithm has a prediction accuracy of 90.91%, while the Naive Bayes algorithm has an accuracy of 63.64%. The C4.5 Decision Tree algorithm is superior in predicting student graduation compared to Naive Bayes, which means that the C4.5 Decision Tree is more effective in identifying students who are likely to pass or not pass. The implementation of the C4.5 Decision Tree algorithm also helps schools make better decisions to support students who require additional attention. The study concluded that the Decision Tree C4.5 algorithm is recommended for use in predicting student graduation because it provides higher accuracy. The results of this research can be used by schools to improve the efficiency of the graduation prediction process and develop more effective and efficient learning programs. Using the right algorithms, schools can be more proactive in identifying students who need additional support, which can reduce academic failure rates and improve the overall quality of education.

Keywords: *Decision Tree, Naive Bayes, YPK Imanuel Nabire Junior High School, Central Papua*

**Corresponding Author: Kristia Yuliawan*

1. INTRODUCTION

Student graduation is a crucial aspect of educational development, as it reflects the success of academic institutions in supporting students' academic journeys [1]. In Nabire Regency, Indonesia, the YPK Immanuel Junior High School faces the challenge of optimizing student graduation rates. Accurate prediction of student graduation can provide valuable insights for educational institutions to implement effective remediation and retention policies [2]. Predicting student graduation can lead to more efficient resource allocation and targeted interventions, ultimately improving the overall quality of education [3]. The implementation of data mining in the educational sector has brought significant benefits[4] . Techniques such as the Naïve Bayes algorithm can be used to predict student graduation by generating models that can identify students at risk of not graduating on time[5]. Machine learning algorithms, such as Decision Tree and Naïve Bayes, have been widely adopted in educational data mining to predict student graduation. These algorithms have their respective advantages, making them suitable for different types of educational data[6].

This study aims to compare the effectiveness of the Decision Tree and Naïve Bayes algorithms in predicting student graduation at YPK Immanuel Junior High School, Nabire Regency. To achieve this objective, the researchers collected data on student academic performance, including attendance, test scores, and final grades, from the school's records. By analyzing the performance of these two algorithms, the study will provide valuable insights to the school's administration on the most effective approach to identify and support students at risk of not graduating

on time, ultimately contributing to the improvement of the overall quality of education. The Decision Tree algorithm is a widely used machine learning technique that constructs a hierarchical tree-like structure to make decisions based on a series of rules . This algorithm is known for its ability to classify data objectively and present results in an easily understandable manner, making it a popular choice among decision-makers[8] . On the other hand, the Naïve Bayes algorithm is a probabilistic classification technique that assumes independence between features in the dataset . Despite this simplistic assumption, the Naïve Bayes algorithm has been shown to perform well in classification tasks, even when the assumption of independence is violated[9] . The Decision Tree and Naïve Bayes algorithms have been successfully applied in the field of educational data mining for predicting student graduation . These algorithms can effectively handle the complex and multifaceted nature of student academic performance data, which often includes factors such as attendance, test scores, and final grades[10] . For the experimental setup and data collection process, the researchers gathered student academic data from the YPK Immanuel Junior High School in Nabire Regency, Indonesia. Previous studies have demonstrated the effectiveness of the Decision Tree and Naïve Bayes algorithms in predicting student graduation . For instance, a study conducted in the Philippines utilized the Naïve Bayes algorithm to predict student graduation by generating a model that could identify students at risk of not graduating on time[11] . Another study compared the performance of the Decision Tree and Naïve Bayes algorithms in predicting study period, using parameters such as college entrance, grades, and secondary school type[12] . Additionally, a research paper analyzed the use of the Decision Tree method with the application of genetic algorithms to predict student graduation accuracy [13].

The researchers will develop two student graduation prediction models using the Decision Tree and Naïve Bayes algorithms. The Decision Tree model will involve constructing a hierarchical tree-like structure based on the student academic performance data, allowing the identification of key factors that influence graduation[14] . The Naïve Bayes model, on the other hand, will utilize a probabilistic approach to classify students based on the assumption of independence between the academic performance variables[15]. Based on the findings of a comparative analysis between the Decision Tree and Naïve Bayes algorithms, it provides recommendations to school administrations and local education offices in Nabire Regency on how to improve the quality of education and support student success. These recommendations may include identifying the most influential factors that contribute to a student's graduation, such as attendance, test scores, or socioeconomic background, and developing targeted interventions to address these factors. Furthermore, it suggests ways to incorporate the predictive model developed in this study into the school's decision-making process, allowing for early identification of at-risk students and the implementation of tailored support programs.

2. RESEARCH METHOD

Research methods are scientific methods to obtain data for specific purposes and uses. When conducting research, we need to follow the applicable rules or rules so that the research results obtained can be considered valid. The research method systematically involves direct observation in the field, and the data collection obtained is then analyzed using statistical techniques to find correlations between the variables studied. In this study, the two algorithms were compared as the best algorithms. The analysis was ranked to predict the graduation of YPK Immanuel Junior High School students in Nabire Regency, Central Papua, using the decision tree method and Naïve Bayes by prioritizing the orange tools. The flow of this study is shown in figure 1.

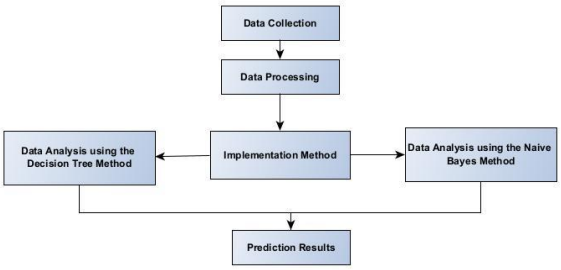

Figure 1. Research Flow

2.1 Data Collection

Data collection is carried out to obtain the necessary information or materials because data are needed to support the prediction of student graduation and learning materials. Collect historical data on YPK Immanuel Junior High School students in Nabire Regency, including academic grades (exam scores). The data collection steps were as follows:

a. Field Study

The author carried out a field study at YPK Immanuel Junior High School, Nabire Regency, for approximately one month. From July 12 to August 15, 2023, in this field study, the author was placed in the computer LAP room and taught ICT subjects in grades XIII & IX[10], so that the author could understand how to teach and know the problems that exist in the school, namely; One of them is that teachers experience obstacles to see the quality of student graduation

b. Interview

The method of collecting data through interviews is carried out through a direct and structured dialogue process between the researcher and the data source, or resource person, namely the principal of the school, P. Yetty Korowa. The purpose of this interview is to take student data that will predict graduation and data taken, namely, in the class of 2022 & 2023 class IXA,

40 students as training data and class IXA students of the class of 2024 are 11 students as test data

c. Literature studies

This method was used to find relevant summaries or facts needed in this study, which were mainly sourced from reference book readings. Articles in previously published research journals. The purpose is to collect data that will be used in the development of the system and all aspects related to the decision tree and Naive Bayes algorithms.

2.2 Data Processing

At this stage, process the data that have been collected so that rules are formed that will be decisions. Before the data in orange connect to the decision tree, the collected data are first cleaned . Data cleaning is carried out to determine which data are used because, at the beginning, there are many attributes so that attributes that are not needed and not processed in the decision tree in this study. The author cleaned the data, such as in the data there is NIM, parents' names, addresses and several other attributes are deleted because these attributes are not needed in this study.

Data processing includes cleaning missing or inconsistent values. By normalizing numerical data by converting categorical data into numerical data, this effort was made to improve and prepare for processing. It considers data collection, data processing, and data cleansing unnecessary in predicting student graduation. The collected data were divided into two sets: data training and data testing. The data mining process uses the C4.5 and Naïve Bayes decision tree algorithms to predict student graduation with high accuracy.

2.3 Algorithm Implementation

The next stage of this study is to implement the algorithm. The implementation of this algorithm is carried out with orange software, connecting the decision tree and naïve Bayes tools to prediction using data testing, which analyzes the data using the decision tree and naïve Bayes algorithms.

a. Data Analysis Using The Decision Tree Method

At this stage, the data analysis was performed using the decision tree algorithm. In general, the stages of the decision tree algorithm involve building a decision tree by selecting attributes as the root and then branches for each value. The decision tree algorithm is a supervised learning algorithm used to create classification models. This algorithm uses the attributes listed in the data to form a tree, where each mode in the tree is a decision based on the value of a certain attribute. The decision tree algorithm can be used to predict student graduation using attributes, such as name, subject grade, grade point average, and graduation status.

In general, the decision tree algorithm to build a decision tree involves several stages, namely, preprocessing data training. The training data are taken from historical data that have been grouped into certain classes, after preparing the next data to determine the root of the tree to calculate the highest gain value of each attribute or based on the lowest entropy index value. Previously, the entropy index value was calculated first with the formula

$$\mathrm{I(S)} \sum_{i=1}^{n} = \frac{|Si|}{|S|} \log_2 \frac{|Si|}{|S|} \quad (1)$$

where S is the set of cases, Si is partition i, n is the number of partitions, and |S| is the number of cases in S. After calculating the entropy value, the gain value is calculated using the following formula:

$$\mathrm{Gain} = \mathrm{I(S)} - \sum_{i=1}^{n} = \frac{|Si|}{|S|}\ l(Si) \quad (2)$$

Information :
S: case set
A: Attribute
n: number of partitions attribute
A |Si |: number of cases in the ith partition
|S|: number of cases in Before

In analyzing this decision tree, it will analyze what attributes are needed in the decision tree algorithm, so that according to the needs, data is prepared and what level of accuracy is produced in predicting student graduation using the decision tree algorithm, which can effectively predict student graduation and help the initial identification of students who are likely not to pass, so as to allow appropriate remediation and retention policies. The C4.5 Decision Tree algorithm accurately predicts student graduation rates with a high degree of accuracy, thus helping schools to develop policies and reduce barriers.

b. Data Analysis Using The Naïve Bayes Method

Analyzing the data at this stage is the same as the data used in data analysis using a decision tree algorithm. The naive Bayes algorithm is a probability-based data-processing method. The Naive Bayes algorithm can be used as a supporting algorithm in the process of creating a decision-tree classification model. The naïve Bayes algorithm can be used to help select attributes that are important in the process of creating a decision tree. The Nave Bayes algorithm can effectively predict student graduation, so it will analyze the level of accuracy generated from the decision tree algorithm to help identify students who are likely not to graduate early, thus enabling appropriate remediation and retention policies. In general, the formula commonly used to calculate the probability of graduation events based on attribute attributes is as follows:

$$P(X|H) = \frac{P(X|H) - P(H)}{P(X)} \quad (3)$$

Information:

X : Data with unknown classes

Q : The X data hypothesis is a specific class

P(H|X) : Probability of hypothesis H based on condition X (posteriori probability)

P(H) : Probabilitas hipotesis H (posteriori probability)

P(X|H) : Probability X based on the conditions in hypothesis H

P(X) : Probability X

2.4 Prediction Results

At this stage, we will explain the results of predicting student graduation with high accuracy using each of the decision tree and naïve Bayes algorithms to compare the two algorithms to determine the best algorithm for predicting student graduation.

3. RESULT AND DISCUSSION

The image shows the user interface (UI) of the Orange Data Mining Software. The relationships between the widgets in figure 2 show the process of building and evaluating the decision tree and Naive Bayes classification models. Training data were used to build the model and test data were used to evaluate the model's performance. The prediction results are displayed in various formats, including decision trees and prediction tables.

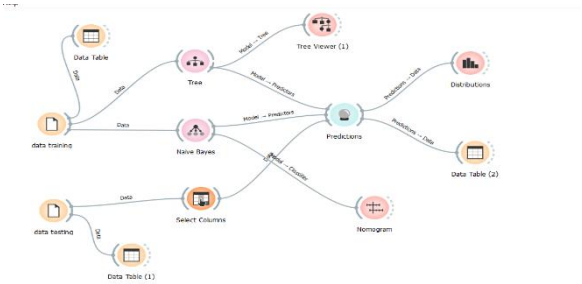

Figure 2. user interface (UI) of the Orange Data

3.1 Data analysis using the Decision tree method

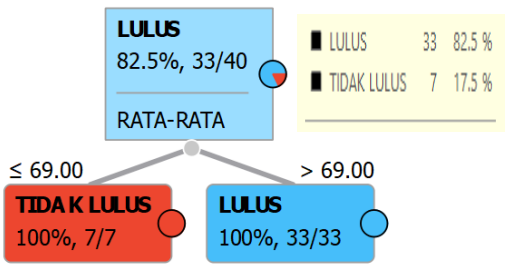

Figure 3: Decision tree model

Based on the analysis of the decision tree shown in figure 3, the exam results show that most of the participants had successfully passed. The second analysis showed that all participants who took the exam, with a minimum score of 69, successfully passed. This shows that the graduation standards set are quite high, and only participants who really master the exam material can pass. The analysis of passing and not passing in a total of 40 participants showed that the number of participants who passed was 33 people with a passing percentage of 82.5%, and the number of participants who did not pass as many as seven people with a percentage of unsuccessful passing of 17.5%. This analysis showed that most participants successfully passed the exam, and only a few participants did not pass the exam.

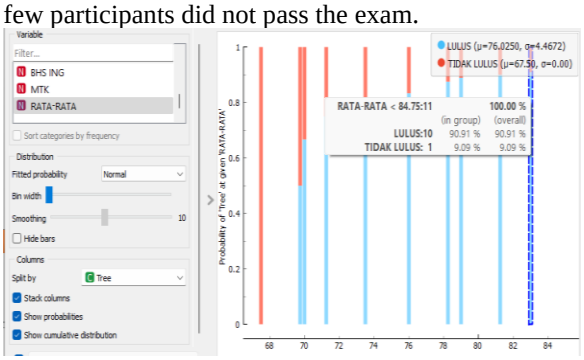

Figure 4. Graph of data analysis results with Decision Tree

Based on figure 4, the results of the analysis using the decision tree algorithm based on the training data are predicted to have an accuracy level of 90.91% and an error value of 9.09%, which shows that the prediction of class IXA students who will graduate in 2024 is 10 students, and one student is predicted not to graduate. The red color in the graph indicates that students have the potential to not pass the average score of the attributes. On the other hand, blue indicates that students are predicted to have the potential to pass the average score in attributes.

3.2 Data analysis using the Naïve Bayes method

Data analysis in the Naive Bayes Algorithm is based on probability and statistics to predict the class of data based on previous experiences. In this analysis, the Naive Bayes algorithm was used to predict student graduation based on students' academic data. The data used in this analysis are data on UN students consisting of academic scores for mathematics, Indonesian, English, science, and average scores from each subject. Based on figure 5, the results of the analysis of Naïve Bayes' student predictions built with training data using the Naïve Bayes algorithm explain that if the average score is $\leq$ 72.5, then the student is predicted not to pass, and if it is $\geq$ 72.5, then the student is predicted to pass. The results of the probability value obtained from the Nomogram are 81% and the error value is 19%

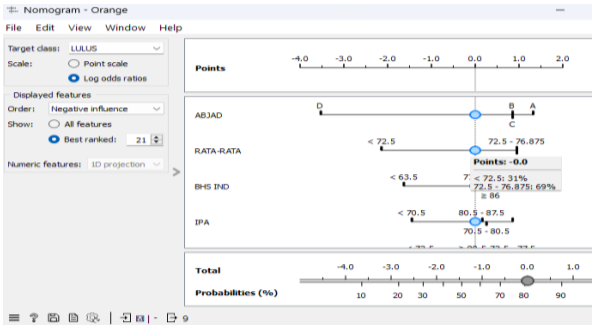

Figure 5. Prediction Analysis with Naïve Bayes' Algorithm

Figure 6 shows the nomogram used to predict the probability of student graduation based on average scores. The results of the analysis using the Naïve Bayes algorithm based on the training data that have been built are predicted to have an accuracy level of 63.64% and an error value of 36%, which shows that the prediction of class IXA students who will graduate in 2024 is seven students, and one student is predicted not to pass. In the above graph, the red color indicates that students have the potential to not pass the grades in the attributes. On the other hand, blue indicates that students are predicted to have the potential to pass the average score in attributes. Thus, the Naive Bayes algorithm can predict student graduation with a relatively high level of accuracy. However, it is important to note that this level of accuracy can vary depending on the quality of the data and complexity of the problem.

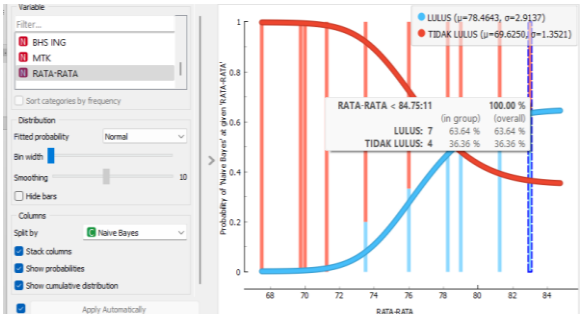

Figure 6. Prediction Analysis Based on Means on the Naïve Bayes Method

3.3 Prediction Results

The results of student graduation predictions using the decision tree and Naïve Bayes obtained different accuracies from the two methods as shown in the Table 1.

Table 1 . Comparison of Algorithm Accuracy Results

No	Method	Number of Passed	Number of Non Grauates	Accuracy Values	Error Values
1	Decision Tree	10	1	90.91%	9.09%
2	Naive Bayes	7	4	63.64%	36.36%

Based on Table 1. shows that decision trees have high accuracy compared to Naïve Bayes, and the accuracy is the degree of proximity between the predicted value and the actual value. In this study, the performance of the decision tree has a higher accuracy of 90.91%, and Naïve Bayes has an accuracy of 63.64% based on the available data. However, both methods were used to predict student graduation in the visualization of the results, which can be helpful in understanding the process of predicting student graduation.

4. CONCLUSION

This study shows that the Decision Tree C4.5 algorithm is superior in predicting student graduation at YPK Immanuel Nabire Junior High School with an accuracy of 90.91% compared to the Naive Bayes algorithm which has an accuracy of 63.64%. This is because the C4.5 Decision Tree algorithm is able to handle variables that interact in a complex and non-linear manner. The C4.5 Decision Tree algorithm makes decisions based on repeated separations from data into sub-groups based on the most informative attributes, so that it can more effectively capture more complex patterns in the data. instead, Naive Bayes' algorithm assumes that each feature or attribute of the data is independent of each other, which often does not correspond to the reality in which the features may interact with each other. This assumption of independence can lead to a decrease in performance, especially in situations where there is a correlation between attributes that affect the prediction results. Therefore, in the context of student graduation prediction involving many interrelated factors, the C4.5 algorithm can provide better performance compared to Naive Bayes. The application of this algorithm uses Orange software, which makes it easy to visualize and analyze the data. The results of the study recommend the use of the C4.5 Decision Tree algorithm because it provides higher prediction accuracy, which can support the development of more effective and efficient learning programs in schools.

5. REFERENCES

[1] A. C. Lagman et al., "Embedding naïve Bayes algorithm data model in predicting student graduation," *in Proceedings of the 3rd international conference on telecommunications and communication engineering*, 2019, pp. 51–56.

[2] A. Gopalakrishnan, R. Kased, H. Yang, M. B. Love, C. Graterol, and A. Shada, "A multifaceted data mining approach to understanding what factors lead college students to persist and graduate," *in 2017 Computing Conference*, 2017, pp. 372–381.

[3] N. Pandiangan, M. L. C. Buono, and S. H. D. Loppies, "Implementation of Decision Tree and Naïve Bayes Classification Method for Predicting Study Period," *in Journal of Physics*: Conference Series, 2020, p. 22022.

[4] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *Ieee Access*, vol. 8, pp. 75264–75278, 2020.
[5] M. T. Sembiring and R. H. Tambunan, "Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)," *in IOP Conference Series: Materials Science and Engineering*, 2021, p. 12069.
[6] M. A. Jassim, "Analysis of the performance of the main algorithms for educational data mining: a review," *in IOP conference series: materials science and engineering*, 2021, p. 12084.
[7] K. Kusrini, A. B. Prasetio, and others, "Prediction of Student Graduation with Naive Bayes Algorithm," *in 2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1–5.
[8] H. S. El-Ghety, I. Emam, and A. M. Ali, "Performance Evaluation of Different Supervised Machine Learning Algorithms in Predicting Linear Accelerator Multileaf Collimator Positioning's Accuracy Problem," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, 2022.
[9] R. Achmad and A. S. Girsang, "Implementation of naive bayes classifier algorithm in classification of civil servants," *in Journal of Physics: Conference Series, 2020*, p. 12018.
[10] A. Purwinarko, W. Hardyanto, and N. P. Aryani, "Academic achievement analysis of Universitas Negeri Semarang students using the naive bayes classifier algorithm," *in Journal of Physics*: Conference Series, 2021, p. 42130.
[11] A. Marzuqi, K. A. Laksitowening, and I. Asror, "Temporal Prediction on Students Graduation using Naïve Bayes and K-Nearest Neighbor Algorithm," *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 682–686, 2021.
[12] N. Renaningtias, J. E. Suseno, and R. Gernowo, "Hybrid Decision Tree and Naïve Bayes Classifier for Predicting Study Period and Predicate of Student's Graduation," *Int. J. Comput. Appl.*, vol. 975, p. 8887, 2018.
[13] A. Maulana, "Prediction of student graduation accuracy using decision tree with application of genetic algorithms," *in IOP Conference Series: Materials Science and Engineering*, 2021, p. 12055.
[14] E. Alyahyan and D. Dü\csteaör, "Decision Trees for Very Early Prediction of Student's Achievement," *in 2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, 2020, pp. 1–7.
[15] P. Subhash and N. Choudhary, "Predicting Instructor Performance using Naive Bayes Classification Algorithm in Data Mining Technique [J]," *Int. J. Comput. Appl.*, vol. 179, no. 22, pp. 9–12, 2018.