1258-Article Text-7356-1-15-20250205 (3).docx

by 1 1

Submission date: 14-Apr-2025 03:56PM (UTC+0700)
Submission ID: 2645562630
File name: 1258-Article_Text-7356-1-15-20250205_3_.docx (346.9K)
Word count: 5215
Character count: 30570

Optimization of Naive Bayes and Decision Tree Algorithms through the Application of Bagging and Adaboost Techniques for Predicting Student Study Success

Endi Febriyanto[1], Wasilah[2*]

[1,2] Institute of Informatics and Business Darmajaya, Lampung, 35142, Indonesia
*wasilah@darmajaya.ac.id

Received: Month xx, xxxx; Revised: Month xx, xxxx; Accepted: Month xx, xxxx.

Abstract: The percentage of student failure in learning still shows a fairly high number. This condition is experienced by many countries, including Ghana, Nigeria, Enugu and Indonesia. Internal factors and external factors of students that vary greatly have the potential to be the cause of failure. This condition cannot be allowed to continue. A special analysis is needed regarding factors that can help improve student grades. Predictions of student success are urgently needed. These predictions can anticipate negative impacts that occur, including increased risk of dropout, decreased student motivation to learn, and individual potential. The Naive Bayes and Decision Tree algorithms have been used to predict student success. However, among its advantages, these two algorithms still have several weakness. It can cause the algorithm's performance not to be as expected. Several methods in ensemble techniques can improve algorithm performance. Two methods that are often used are Bagging and Adaboost. Bagging and Adaboost can help improve the performance of classification algorithms. This Study will combine Bagging and Adaboost into the Decision Tree and Naïve Bayes algorithms to optimize the results in predicting student success. The stages carried out are data collection, data pre-processing, data split,data processing and evaluation model. The results show that Bagging and Adaboost techniques have been proven effective in improving accuracy, precision, recall, and F1-Score performance. Combining the naïve Bayes algorithm with Adaboost increases accuracy, precision and F1 score significantly by 1.95%, 28.98%, and 15.79%.

Keywords: Naive Bayes Optimization, Decision Tree Optimization, Bagging, Adaboost, Student learning success

1 Introduction

The world of education should be the focus of government attention in order to improve the quality of human resources. Formal schools are an education system that has been optimally regulated, so that it is expected to be able to create human resources that are able to advance the nation.[1]. Predicting student success plays a vital role in improving the effectiveness of education. By understanding the factors that influence success, educational institutions can provide appropriate interventions to students in need. This can reduce dropout rates, increase student retention, and provide additional support to students at risk of dropping out. Accurate predictions also allow for the development of individualized learning programs, helping students better reach their academic potential, thereby improving the overall quality of education.

The percentage of student failure in learning shows a fairly high number. This condition is experienced by many countries, including Ghana, Nigeria, Enugu and Indonesia. WAEC reports consistently show

unsatisfactory student performance in the Ashanti Region of Ghana in biology [10][12]. Stakeholders in Nigeria have raised concerns over the trend of poor academic performance of students in all categories of schools Nigeria. In addition, the examination of students' academic performance in Basic Science in Enugu State, as indicated by the results of the Basic Education Certificate Examination (BECE) from 2018 to 2022, shows an alarming pattern of below average achievement [2]. Many countries experience similar conditions including Indonesia.

This condition cannot be allowed. A special analysis is needed regarding factors that can help improve student grades. The inability to predict student study success can have a significant negative impact. Without accurate predictions, educational institutions cannot identify student abilities, resulting in a lack of appropriate intervention. This condition can increase the risk of dropout, reduce student learning motivation, and cause loss of individual potential. Predicting student study success plays an important role in improving the effectiveness of education. Many factors influence the success of a study. Traditional approaches are often not effective enough. Machine learning approach has many algorithms that can be used to solve prediction problems, including the Naive Bayes and Decision Tree algorithms. The machine learning approach can be used to analyze complex data more deeply and accurately.

Naive Bayes and Decision Tree algorithms have been used in an attempt to predict students' study success.[2] [3]. Decision Tree has the advantage of being flexible so that it can improve the quality of the decisions produce[4]. Naive Bayes, the simplicity of its calculations, thus allowing for faster and more efficient processes. The naïve Bayes method only requires a relatively small amount of training data to determine the parameters needed in the classification process[5]. However, both algorithms have weaknesses. Decision Tree is prone to overfitting and unstable decisions, while the naïve Bayes method only supports attributes with discrete or discretized data types, and does not support attributes with continuous (numeric) values, so that all attributes become independent. In addition, these attributes can contribute to the predicted attributes[6].

There are two methods that are often used in ensemble techniques, namely bagging and boosting[7][8]. Bagging and boosting have the ability to support unstable classification algorithm[9]. The formulation of the problem in this study is: How is the application of Bagging and Adaboost (adaptive boosting) techniques on the Decision Tree and Naïve Bayes algorithms for predicting student success? and what are the results of the analysis of the best algorithm performance in predicting student success?.

Several studies have been conducted on the use of bagging and Adaboost methods to improve the performance of Decision Tree and Naïve Bayes algorithms. Research using this method has been conducted to predict the success of studies including the success of students at universities[10][11], predicting success in Distance education[12]. In this study, what will be done is to compare the performance of the Decision Tree and Naïve Bayes classification algorithms with the application of ensemble techniques, namely the Bagging and Adaboost techniques to predict student learning success. Furthermore, a comparative analysis of the results of the application of the algorithm will be carried out to determine the performance of the Decision Tree algorithm, Naïve Bayes combined with ensemble techniques in making predictions.

Ensemble techniques have varying performance depending on the characteristics of the dataset and the algorithm used. Several studies have compared the performance of Bagging and Adaboost using various machine learning algorithms[13]. The use of ensemble techniques such as Bagging and Adaboost has been proven effective in improving the performance of classification models. Studies have shown that by combining several classification models (such as Naive Bayes and Decision Tree) using these ensemble techniques, more accurate predictions can be produced[14]. There is research using this method at universities[15][11], **but still very few focus on the classification of student learning outcomes**[16]. Generally research focuses on distance education[12], classifying learning types in education, learning styles[17], and student performance[18]. Research on predicting academic success has been conducted using various data, including those related to learning styles, gender, age, personality, academic background, and others. **However, it has not been focused on academic grades in detail.** While academic grades are a factor that is very attached to students and has a high potential to have a high influence. **In this study, an ensemble technique will be combined with the naïve Bayes algorithm and decision tree, using data that focuses on the details of students' academic grades and extracurricular activities. This is a novelty that is expected to improve algorithm performance.**

2 Literature Review

Decision trees are a popular and powerful tool in data science and machine learning, used for classification, regression, and other predictive modeling tasks. They are valued for their simplicity, interpretability, and ability to handle both categorical and continuous data. Decision trees work by recursively partitioning the data into subsets based on the value of input features, creating a tree-like model of decisions. The process starts with a root node and splits the data at each node based on certain criteria until reaching leaf nodes, which represent the final decision or classification [19]. The construction of decision trees involves selecting the best attribute to split the data at each node, which can be done using various measures such as Gain Ratio, Gini Index, or other node splitting measures [20].

Naïve Bayes is a popular algorithm used for text classification and sentiment analysis due to its simplicity and efficiency. It operates under the assumption of attribute independence, which can sometimes be violated in real-world data. This analysis explores how well Naïve Bayes models reflect people's views by examining their performance in various applications [21].

Ensemble methods, such as boosting and bagging, are powerful techniques used to improve the accuracy of classifiers by combining multiple models. Among these, AdaBoost is particularly well-known for its effectiveness in binary classification tasks. Boosting: This method sequentially trains classifiers, with each new classifier focusing on the errors made by the previous ones. AdaBoost is a popular boosting algorithm that adjusts the weights of misclassified instances, making it highly effective for binary classification tasks. Bagging, or Bootstrap Aggregating, involves training multiple classifiers independently on different subsets of the training data, created through bootstrapping. The final prediction is made by averaging the predictions of all classifiers, which helps in reducing variance and preventing overfitting [7].

Research on the classification of student graduation using data mining techniques has been widely conducted. Research [18] using the Naïve Bayes algorithm showed quite good performance, namely: 87% accuracy, 91% precision for the First Class class, 78% recall, and 81% F1-score. The features used in this study include department, level, weekly study time, satisfaction with the learning system, engagement in group discussions, engagement in school policy, attendance rate, engagement in curricular activities, learning method, accommodation type, gender, and age.

Research [15] used the Stacking Ensemble method to predict student graduation, resulting in 95% accuracy, precision, recall, and F1-score in the range of 91%-93%. Meanwhile, other ensemble methods such as Bagging and Boosting generally have accuracy, precision, recall, and F1-score in the range of 68%-73%. The features used in this study were assignment, quiz, mid-test, and final test. Research [22] combined the Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) algorithms and showed a high level of accuracy in predicting students at risk of failure. This model achieved excellent evaluation metrics, with accuracy, precision, recall, and F1-score exceeding 90%. The features used include Lecture Notes, Materials, Video, Live Attendance, and Live Activities. In addition, research [23] comparing Decision Tree (DT) and Artificial Neural Network (ANN) showed similar performance, with an accuracy of 62.3%, precision of 36.7%, recall of 91.7%, and F1-score of 52.4%. The features used in this study include semester level, gender, nationality, birthday place, resources visit, announcement response, and extra discussion.

The performance evaluation metrics used include accuracy, precision, recall, and F1 Score. Accuracy provides an overview of how well a model performs in classifying data. Percentage of correct predictions out of total predictions. Here is the formula for calculating the accuracy value:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Precision is the ratio between the data correctly classified as true positives divided by the total number of classified data. It is the Percentage of correct positive predictions out of the total positive predictions made by the model. Precision provides information about how reliable a model is in classifying data as positive. The higher the precision value, the fewer negative cases are incorrectly identified as positive, indicating that the model is more likely to produce true positive results. The formula used is:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall, also known as sensitivity, is a measure that indicates how well a classification model can identify actual positive cases in a dataset. In classification, recall is calculated as the ratio of correctly predicted positives (true positives) divided by the total number of true positive cases (true positives + false negatives). Percentage of correct positive predictions out of total actual positive cases in the data. In other words, recall measures the ability of a model to "remember" or "detect" all existing positive cases without missing any. Here is the recall formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 Score is an evaluation metric that measures the balance between precision and recall. Here is the formula of F1 Score:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{RecalL}}{\text{Precision} + \text{Recall}} \quad (4)$$

3. Methodology

The identification process begins with a literature study. A literature study is conducted to identify attributes that have the potential to become criteria. At this stage, it is a systematic process of reviewing and analyzing literature, publications, scientific articles, books, and other relevant information sources. Criteria identification is based on the results of literature studies and expert confirmation. In this process, semi-structured interviews are conducted with experts. Interviews are based on questions that refer to the results of the literature review. The stages in this Study are data collection, data pre-processing, data split, data processing and evaluation model [15]. These stages are shown in Figure 1.

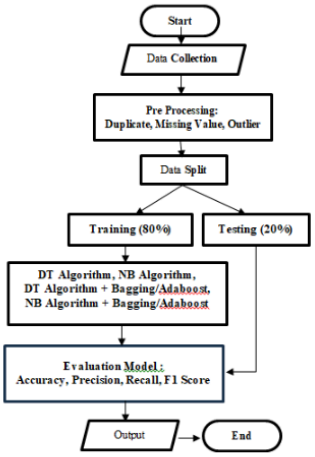

Figure 1: The stages of Research

3.1. Data Collection

This research uses data sourced from Kaggle. The link address is https://www.kaggle.com/datasets/mexwell/student-scores.

3.2. Data Pre-processing

In data mining, the preprocessing stage plays a very important role [24][25]. The main stage is the cleaning stage[26]. The steps for cleaning data are deleting irrelevant columns, checking for missing data, checking for duplicate data, checking for outlier values , and checking for categorical data consistency. Data Preprocessing uses Google Collaboration

3.3. Evaluation model

In this study, the data processing techniques used are the Decision Tree and Naïve Bayes algorithms. The ensemble techniques that will be used are Bagging and AdaBoost. This process involves dividing the classification into six different scenarios. The scenarios that will be carried out are the Implementation of Naive Bayes (NB) and Decision Tree (DT) algorithms, the Application of the Bagging Technique on Naive Bayes (NB+BG) and Decision Tree (DT+BG), the Implementation of Adaboost on Naive Bayes (NB+ADB) and Decision Tree (DT+ADB). Data processing using the Rapid Miner application.

The result analysis stage focuses on comparing results. This stage is carried out based on the test results in six scenarios. These six scenarios include testing the decision tree algorithm, naïve Bayes algorithm, Decision Tree + Bagging algorithm, Decision Tree + Adaboost algorithm, naïve Bayes + Bagging algorithm, and naïve Bayes + Adaboost algorithm. We can determine which method significantly improves classification performance by comparing these values.

4. Results

4.1. Data Collection

Based on the Study's results, several factors are related to student success. The factor are academic performance, demographic factors, school and environmental context, family and social environment, student engagement, and institutional policies. [27][28][29]. The results are the basis for confirmation to the expert to validate the suitability of the attributes to be used. The experts involved in this Study were senior teachers and principals. The experts have master's qualifications in education and are experienced in managing senior high schools. Some of the points produced are academic and non academic factors. Academic factors related to score of subject and non academic factors include extracurricular activities.

In this Study, the dataset used is the student scores dataset downloaded from the Kaggle.com. The amount of initial data before the preprocessing process is 2000 data. This dataset consists of 19 attributes, 18 of which function as predictor variables, including ID first name, last name, email, gender, absence day, extracurricular activities, weekly self-study hours, career aspiration, math score, history score, physics score, chemistry score, biology score, geography score. Meanwhile, academic success is another attribute that acts as a target variable or label. This target variable has two output values, namely pass and fail.

4.2. Data Pre-processing

4.2.1.Remove irrelevant columns.

This process aims to produce a cleaner, more consistent, and more appropriate dataset. It does this by deleting attribute columns that do not affect the modelling process. Some of the deleted columns or attributes are ID, first name, last name, and email. The remaining features are all numeric data types, except part time

job, extracurricular_activities and career aspiration. The features involved are divided into independent variables and dependent variables (labels).

4.2.2. Changing category columns to numeric columns

At this stage, the academic success column is changed to a numeric column to see the relationship between the category and target variables. The numeric values are 1 and 0. Value 1 = 'pass' and value 0 = 'fail'.

4.2.3. Checking for missing value and duplicate data

The data cleaning process is carried out on the student score dataset, which has inconsistent N/A and missing values. Duplicate data checking reduces redundancy, improves model accuracy, and improves data processing efficiency. The results of the missing value process show that no data was found missing, and no duplicate data was found.

4.2.4. Checking for outlier values

This stage is checking for outliers. Outlier checking is performed on the attributes involved. Outliers exist in several variables, including math scores, biology scores, and average values. In this study, outliers are handled in truncation. Values below the lower bound are set to become the lower limit values themselves, and values above the upper bound are changed to upper limit values. The instruction for handling outlier values:

4.3. Confusion matrix

Confusion matrix testing was conducted on six research scenarios. The Confusion matrix results are shown in Table 1.

Table 1: Confusion Matrix

	DT	NB	ADB + DT	ADB + NB	BG + DT	BG + NB
TP	1524	1494	1716	1718	1719	1685
FP	4	5	1	7	3	6
FN	6	36	6	4	3	37
TN	66	65	77	71	75	72

The confusion matrix results show that Bagging + Decision Tree produces the highest number of correct predictions (1719), while Naïve Bayes has the lowest results (1524). The highest false positives are in Adaboost + Naïve Bayes (7). This indicates that this model is more accurate in predicting students who actually graduate. The highest false negatives are in Bagging + Naïve Bayes (FN=37). This indicates that this model is more often wrong in classifying students who actually graduate, the highest true negatives are in Adaboost + Decision Tree.

4.4. Evaluation model

Validation and Testing of the model on student graduation data is carried out to evaluate the model's performance in predicting the possibility of students graduating and failing. Testing is carried out based on attributes and preprocessing data. Testing is carried out on six scenarios, namely: Decision Tree algorithm, Naive Bayes algorithm, combining Bagging Techniques on Naive Bayes and Decision Tree algorithms, combining Adaboost Techniques on Naive Bayes and Decision Tree algorithms.

K-fold Cross-Validation was performed using 10 folds. Table 2 shows the test results for the Decission Tree algorithm and Naïve Bayes Algorithm.

Table 2: Evaluation of Decission Tree and Naïve Bayes Algorithm

Validation	Decission Tree Algorithm	Naïve Bayes Algorithm
Accuracy	99,38%	97,44%
Precision	92,50%	66,16%
Recall	94,29%	92,86%
F1 Score	93,36%	77,27%

The next scenario are implementation of the Bagging Technique on Decision Tree (DT+BG) and Adaboost on Decision Tree (DT+ADB). Table 3 shows the test results.

Table 3: Evaluation of Decission Tree +BG and Decission Tree + ADB Algorithm

Validation	Decission Tree + BG Algorithm	Decission Tree + ADB Algorithm
Accuracy	99,67%	99,61%
Precision	96,39%	93,28%
Recall	96,07%	98,75%
F1 Score	96,23%	97,55%

The next scenario are implementation of the Bagging Technique on Naive Bayes (NB+BG) and Adaboost on Naive Bayes (NB+ADB). Table 4 shows the test results.

Table 4: Evaluation of Naïve Bayes +BG and Naïve Bayes + ADB Algorithm

Validation	Naïve Bayes + BG Algorithm	Naïve Bayes + ADB Algorithm
Accuracy	97,61%	99,39%
Precision	68,17%	95,14%
Recall	92,32%	91,07%
F1 Score	78,35%	93,06%

4.5. Feature Importance

Feature importance testing is conducted to see the influence between attributes. Testing is conducted on six research scenarios. The test results on each decision tree algorithm, Naïve Bayes and decision tree, Naïve Bayes with the addition of Adabost and bagging are shown in Figure 2.

Figure 2: Feature importance

The test results show that the most influential features in sequence are the three highest math_score, history_score, English_score. While the feature with the least influence is extraculicular_activities. In all six scenarios, the level of importance in model testing is the same.

5. Discussions

Based on the table above, the results of the comparative Testing on the student graduation grade dataset using the decision tree algorithm get accuracy value, precision, recall and F1 scores of 99.38%, 92.50%, 94.29% and 93.36%. The naïve Bayes algorithm produces a performance of 97.44%, 66.16%, 92.86% and 77.27%. In the third scenario, the merger decision tree algorithm and bagging obtained 99.67%, 96.39%, 96.07% and 96.23%. It shows an increase in inaccuracy, precision, recall and F1 score, namely 0.29%, 3.89%, 1.78% and 2.87%. In the fourth scenario, the combination of decision and AdaBoost produces an accuracy value, precision, recall and F1 score of 99.61%, 93.28%, 98.75%, and 97.55%. It shows an increase in accuracy, recall, precision and F1 Score: 0.23%, 0.83%, 4.46% and 4.19%. In the fifth scenario, between the naïve Bayes algorithm and the bagging technique, the accuracy value precision, recall, and F1 score are 97.61%, 68.17%, 92.32%, and 78.35%. It shows that the amount is 0.17%, 1.01%, and 1.08%. While recall, there was a decrease of 0.54%. In the sixth scenario, namely the combination of the naïve Bayes algorithm with Adaboost, the accuracy value is precision, recall and F1 score of 99.39%, 95.14%, 91.07% and 93.06%. It shows an increase of 1.95%, 28.98%, and 15.79%. Meanwhile, recall decreased by 1.79%. The performance for each evaluation is shown in graphical form in the Figure 3.

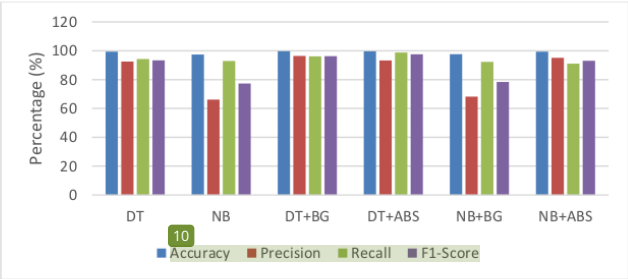

Figure 3: The performance for each evaluation.

Figure 3 shows that, for the first: the Decision Tree algorithm achieved an impressive accuracy rate of 99.38%, indicating exceptionally high performance. The Naive Bayes algorithm demonstrated a lower accuracy of 97.44%. The application of Bagging enhanced the accuracy of the Decision Tree to 99.67%, while the use of Adaboost increased it to 99.61%. For Naive Bayes, the Bagging technique marginally improved accuracy to 97.61%, and with the addition of Adaboost, it reached 99.39%. Overall, both Bagging and Adaboost significantly improved the accuracy of the Decision Tree and Naive Bayes algorithms, with the Decision Tree consistently maintaining superior performance.

The second, Decision Tree algorithm, without ensemble techniques, exhibited a precision rate of 92.50%. In contrast, Naive Bayes displayed a considerably lower precision of 66.16%. However, Naive Bayes showed significant improvement when combined with Bagging and Adaboost, achieving precision rates of 68.17% and 95.14%, respectively. When combined with Bagging, the Decision Tree attained the highest precision rate of 96.39%, while the combination with Adaboost yielded a slightly lower precision of 93.28%.

The third, Decision Tree algorithm demonstrated a recall of 94.29%, reflecting good sensitivity, whereas Naive Bayes had a slightly lower recall rate of 92.86%. The application of Adaboost to the Decision Tree significantly increased the recall to 98.75%. In contrast, combining Bagging and Adaboost with Naive Bayes resulted in a recall that remained lower than the Decision Tree.

The last, F1-Score for the Decision Tree was 93.36%, while Naive Bayes lagged with a much lower score of 77.27%. The application of Bagging improved the Decision Tree's F1-Score to 96.23%, and the implementation of Adaboost further enhanced it to 97.55%. For Naive Bayes, the Bagging method yielded an F1-Score of 78.35%, and Adaboost succeeded in elevating it to 93.06%.

Some previous studies that have been conducted in predicting learning success include using the Stacking Ensemble, Random Forest (RF) + XGBoost and Decision Tree (DT) and Artificial Neural Network (ANN). The performance produced from previous studies and the performance of this study are shown in Table 4.

Table 4: Previous research performance and research results

	Accuracy	Precision	Recall	F 1 score
Naïve Bayes [18]	**87%**	**91%,**	**78%,**	**81%**
Stacking Ensemble [15]	95%	91%-93%.	91%-93%.	91%-93%.
Random Forest (RF) + XGBoost [12]	90%.	90%.	90%.	90%.
Decision Tree (DT) and Artificial Neural Net(ANN) [23]	62,3%,	36,7%,	91,7%,	52,4%.
Decision Tree (DT)	99,38%	92,50%	94,29%	93,36%
Naïve Bayes (NB)	97,44%	66,16%	92,86%	77,27%
Decision Tree (DT) + BG	99,67%	96,39%	96,07%	96,23%
Naïve Bayes (NB) + BG	97,61%	68,17%	92,32%	78,35%
Decision Tree (DT) + ADB	99,61%	93,28%	98,75%	97,55%
Naïve Bayes (NB) + ADB	99,39%	95,14%	91,07%	93,06%

Performance comparison between previous research and the results of this study shows a significant increase. The highest result for accuracy is Adaboost + Naïve Bayes, the high precision is Bagging + Decision Tree and Recall and F1 score on Decision Tree (DT) + ADB. Likewise, the results of the confusion matrix test conducted showed that the selection of algorithms and ensemble techniques had a significant impact on the balance of precision and recall in predicting student study success. Bagging + Decision Tree is better at reducing false negatives, so it can help institutions identify students at risk of failing more accurately. In contrast, Adaboost + Naïve Bayes is useful for ensuring that students who are predicted to graduate actually graduate. Institutions can utilize these findings in academic monitoring systems, for example, by adjusting tutoring interventions or academic support strategies based on the prediction error patterns of the models used.

These findings underscore that ensemble techniques, particularly Bagging in Naive Bayes, can effectively detect students at risk of academic difficulties, which is crucial for educational settings. The high sensitivity of these models allows early identification of students who may be struggling, enabling timely interventions.

Educational institutions can leverage these optimized predictive models to design more precise intervention programs, such as targeted tutoring, counselling, or career guidance, based on identified risk factors like high or low test scores. Moreover, such models facilitate more efficient and cost-effective intervention planning by focusing on students needing the most support rather than broad and untargeted approaches.

However, this research still has limitations and needs to be developed by combining features more completely and using a larger data set or using appropriate primary data.

6. Conclusions

Decision Tree algorithm, the Bagging and AdaBoost techniques have demonstrated their effectiveness in enhancing machine learning models' performance, accuracy, precision, recall, and F1-Score. The Decision Tree consistently outperformed the Naive Bayes algorithm, especially when combined with Bagging and Adaboost, leading to superior results in both accuracy and precision. Although the Naive Bayes algorithm showed notable improvement when integrated with Adaboost, its performance remained lower than that of the Decision Tree. Among the combinations tested, the Decision Tree coupled with Adaboost emerged as the most optimal in achieving high accuracy, precision, recall, and F1-Score. This finding suggests that even relatively simple algorithms can be highly effective for predicting students' academic success through ensemble methods.

Based on the results of the analysis and conclusions that have been described, the following are suggestions for further research:

- Further research is recommended to explore other algorithms such as Random Forest, Gradient Boosting, or SVM that may provide better results in predicting student study success.
- The use of larger and more varied datasets is also important to increase the validity of the results.
- In addition to using Bagging and Adaboost, other ensemble techniques such as Stacking or XGBoost can be considered to improve model performance.
- Implementing deeper cross-validation and hyper parameter tuning can ensure an optimal model.
  Further research also needs to analyze the factors that influence prediction using feature importance or SHAP values.

Acknowledgments

We express our highest appreciation to the KEMENRISTEK DIKTI which has provided financial support for this research through the Master's Thesis Research Program. We also express our gratitude to the Darmajaya Information and Business Institute for their support so that this research can be finished.

References

[1] M. R. Wisroni Wisroni, "Educational Alternatives to the Empowerment Process Village Community," *Educ. Soci.*, 2022, doi: 10.24036/spektrumpls.v10i4.122040.

[2] D. Pradana and E. Sugiharti, "Implementation Data Mining with Naive Bayes Classifier Method and Laplace Smoothing to Predict Students Learning Results," *Recursive J. Informatics*, vol. 1, no. 1, pp. 1–8, 2023, doi: 10.15294/rji.v1i1.63964.

[3] Y. B. Y. Alsariera, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, 2022, doi: DOI:10.1155/2022/4151487.

[4] Géraldin Nanfack, "Constraint Enforcement on Decision Trees: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, 2022, doi: 10.1145/3506734.

[5] Mahmoud Ragab, Ahmed M. K. Abdel Aal, "Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques," *Comput. Sci. Educ.*, 2021, doi: DOI:10.1155/2021/6241676.

[6] Shenglei Chen, "A novel selective naïve Bayes algorithm," *Shenglei Chen*, vol. 152, 2020, doi: 10.1016/j.knosys.2019.105361.

[7] Sergio González, S. García, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Inf.*

Fusion, 2020, doi: 10.1016/j.inft 2020.07.007.
[8] Y. S. Ibomoiye Domor Mienye, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *Comput. Sci. th.*, 2022, doi: 10.1109/ACCESS.2022.3207287.
[9] A. Plaia, Simona Buscemi, "Comparing Boosting and Bagging for Decision Trees of Rankings," *A. aia, Simona Buscemi*, 2021, doi: 10.1007/s00357-021-09397-2.
[10] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach," *IEEE Access*, vol. 11, no. December, pp. 136091–136 8, 2023, doi: 10.1109/ACCESS.2023.3336987.
[11] Olugbenga Wilson Adejo, "Predicting student academic performance using multi-model heterogeneous ensemble appro ," *J. Appl. Res. High. Educ.*, 2017, doi: 10.1108/JARHE-09-2017-0113.
[12] Halit Karalar, "Predicting students at risk of academic failure using ensemble model during pandemic in a distance learnin ystem," *Int. J. Educ. Technol. High. Educ.*, vol. 18(63), 2021.
[13] A. S. A. Zakaria, "The Best Ensemble Learner of Bagged Tree Algorithm for Student Performance ediction," *Comput. Sci. Educ.*, 2020, doi: 10.3233/faia200552.
[14] C. Jalota, "An Effectual Model for Early Prediction of Academic Performance using Ensemble Classification," no. 02, pp. 19–33, 2023.
[15] NAVEED ANWER BUTT and Z. MAHMOOD1, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach," *Digit. Object Identifier 10.1109/ACCESS.2023.3336987*, 2023.
[16] Wattana Punlumjeak, Sitti Rugtanom, "Improving Classification of Imbalanced Student Dataset Using Ensemble Method of Voting, Bagging, and Adaboost with Under-Sampling Technique," *Comput. Sci.*, 2018, doi: 10.1007/978-981-10-6451-7_41.
[17] L. Lisnawita, "Implementation of Naïve Bayes for Classification of Learning Types," *urnal Teknol. In n Komun.*, 2022.
[18] B. Akanbi, "Application of Naive Bayes to Students' Performance Classification," *Asian J. Probab. Stat.*, vol. 25, 2023, doi: 10.9734/AJPAS/2023/v25i1536.
[19] S. E. Tony Thomas, Athira P. Vijayaraghavan, "Applications of Decision Trees," *Comput. Sci. Math.*, 9, doi: 10.1007/978-981-15-1706-8
[20] Nour El Islem Karabadji, Ilyes Khelf, "A data sampli and attribute selection strategy for improving ision tree construction," *Expert Syst. Appl.*, 2019, doi: 10.1016/J.ESWA.2019.03.052.
[21] N. Normah, "Naïve yes Algorithm For Sentiment Analysis Windows Phone Store Application eviews," *Sink. Sci.*, 2019, doi: 10.33395/SINKRON.V3I2.242.
[22] H. Karalar, C. Kapucu, and H. Gürüler, "Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system," *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, 2021, doi: 10.1186/ 41239-021-00300-y.
[23] S. Hussain *et al.*, "Significance of Education Data Mining in Student's Academic Performance Prediction and Analysis," , vol. 5, no. 3, pp. 215–231, 2023.
[24] M. S. Humaira Jamshed, "Data Preprocessing: A preliminary step for web data mining," *3C Tecnol. innovación Apl. a la pyme*, pp. 206–221, 2019, doi: 10.17993/3CTECNO.2019.
[25] A. F. M. A. Rosid, "The Novel Method for Data Preprocessing CLI," *Adv. Intell. Syst. Technol.*, 2022, doi: 10.53759/aist/978-9914-9946-1-2_21.
[26] Elvin Jafarov, "Data Cleaning Before Uploading to Storage," *Equipment, Technol. Mater.*, 2023, doi: 10.36962/etm13012023-117.
[27] Claire McKinley Yoder, M. Cantrell, "Disparities in High School Graduation by Identity and Disability Using Intermediate and Long-Term Educational Outcomes," *J. Sch. Nurs.*, 2022, doi: 10.1177/10598405221078989.
[28] Guili Zhang, T. Anderson, "Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study," *J. Eng. Educ.*, 2004, doi: 10.1002/j.2168-9830.2004.tb00820.x.
[29] P. K. L. Michaele Aubourg, "The Non-Financial Factors that Can Explain the Low Graduation Rate of Black Students in Higher Education," *Int. J. community Dev. Manag. Stud.*, vol. 7, 2023, doi: 10.31355/93.

1258-Article Text-7356-1-15-20250205 (3).docx

ORIGINALITY REPORT

25%	19%	20%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.frontiersin.org Internet Source	2%
2	pmc.ncbi.nlm.nih.gov Internet Source	1%
3	www.jatit.org Internet Source	1%
4	www.publications.scrs.in Internet Source	1%
5	Hasbanur Hafidz, M. Fakhriza. "Comparison of Naive Bayes Algorithms and Decision Tree for Classifying Hero Fighter Items in the Mobile Legends", Journal of Applied Science, Engineering, Technology, and Education, 2024 Publication	1%
6	repository.polinela.ac.id Internet Source	1%
7	journal.unnes.ac.id Internet Source	1%
8	Submitted to University of Auckland Student Paper	1%
9	journal.50sea.com Internet Source	1%
10	www.mdpi.com Internet Source	1%
11	www.coursehero.com Internet Source	

		1 %
12	api.slingacademy.com Internet Source	<1 %
13	docksci.com Internet Source	<1 %
14	www.springerprofessional.de Internet Source	<1 %
15	repo.itera.ac.id Internet Source	<1 %
16	www.researchgate.net Internet Source	<1 %
17	Submitted to University of Bedfordshire Student Paper	<1 %
18	Nisa Pirsingki, Rizky Wandri. "Sentiment Analysis of #Saverafah Hashtag on TikTok Using Naive Bayes and Decision Tree Methods", Jurnal Informatika, 2025 Publication	<1 %
19	www.geeksforgeeks.org Internet Source	<1 %
20	www.semanticscholar.org Internet Source	<1 %
21	"Transdisciplinary Engineering for Complex Socio-technical Systems – Real-life Applications", IOS Press, 2020 Publication	<1 %
22	sifisheriessciences.com Internet Source	<1 %
23	eprints.umm.ac.id Internet Source	<1 %

24	Nurhadiyanto, Supeno Mardi Susiki Nugroho, Eko Setijadi. "Classification of Aircraft Inspection Result Using K-Nearest Neighbors", 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2019 Publication	<1 %
25	Submitted to University of Sussex Student Paper	<1 %
26	discovery.researcher.life Internet Source	<1 %
27	www.tutorialaicsip.com Internet Source	<1 %
28	journals.plos.org Internet Source	<1 %
29	www.sctce.ac.in Internet Source	<1 %
30	I Gusti Ayu Nandia Lestari, Dewa Gede Hendra Divayana, Kadek Yota Ernada Aryanto. "A Concentration Selection In Study Programs Using SMOTE Techniques With Ensemble Learning Algorithms", 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS), 2023 Publication	<1 %
31	J. F. Torres, S. Valencia, F. Martínez-Álvarez, N. Hoyos. "Chapter 1 Predicting Wildfires intheCaribbean Using Multi-source Satellite Data andDeep Learning", Springer Science and Business Media LLC, 2023 Publication	<1 %
32	cn.overleaf.com Internet Source	<1 %
	ijai.iaescore.com	

33	Internet Source	<1%
34	mdpi-res.com Internet Source	<1%
35	JEANNE KLEYN. "The validity of injecting drug users' self-reports about sexually transmitted diseases: a comparison of survey and serological data", Addiction, 5/1993 Publication	<1%
36	ojs3.unpatti.ac.id Internet Source	<1%
37	"Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)", Springer Science and Business Media LLC, 2020 Publication	<1%
38	Durgesh Kumar Mishra, Nilanjan Dey, Bharat Singh Deora, Amit Joshi. "ICT for Competitive Strategies", CRC Press, 2020 Publication	<1%
39	Yuni Yamasari, Hani Nafisah Amaliya, Rina Harimurti, Andi Iwan Nurhidayat, Ari Kurniawan, Paramitha Nerisafitra. "Classification via Clustering for Subject-based Scientific Fields in Kindergarten Students", 2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE), 2023 Publication	<1%
40	dblp.uni-trier.de Internet Source	<1%
41	napier-repository.worktribe.com Internet Source	<1%

42	Antonella Plaia, Simona Buscemi, Johannes Fürnkranz, Eneldo Loza Mencía. "Comparing Boosting and Bagging for Decision Trees of Rankings", Journal of Classification, 2021 Publication	<1%
43	Shenglei Chen, Geoffrey I. Webb, Linyuan Liu, Xin Ma. "A novel selective naïve Bayes algorithm", Knowledge-Based Systems, 2020 Publication	<1%
44	Woodley, Alan, Chappell, Timothy, Geva, Shlomo, Nayak, Richi. "Efficient feature selection and nearest neighbour search for hyperspectral image classification", 'Institute of Electrical and Electronics Engineers (IEEE)', 2016 Internet Source	<1%
45	dergipark.org.tr Internet Source	<1%
46	Submitted to Auckland University of Technology Student Paper	<1%
47	Kun Fu, Zhen Liu, Xueyou Ren, Shenning Zhang. "Design and research of educational mode in context of teaching gamification", Entertainment Computing, 2024 Publication	<1%
48	Marienel N Velasco, Abegail A. Malabuyoc, Glenn V. dela Cueva, Karina L. Enriquez. "Predicting Licensure Examination Performance Using Data Mining Techniques", 2023 8th International Conference on Business and Industrial Research (ICBIR), 2023 Publication	<1%
	doaj.org	

49	Internet Source	<1%
50	link.springer.com Internet Source	<1%
51	optimizdba.com Internet Source	<1%
52	pearl.plymouth.ac.uk Internet Source	<1%
53	A. R. Mohamed Yousuff, M. Zainulabedin Hasan, R. Anand, M. Rajasekhara Babu. "Leveraging deep learning models for continuous glucose monitoring and prediction in diabetes management: towards enhanced blood sugar control", International Journal of System Assurance Engineering and Management, 2024 Publication	<1%
54	Submitted to Erasmus University of Rotterdam Student Paper	<1%
55	hal.science Internet Source	<1%
56	ouci.dntb.gov.ua Internet Source	<1%
57	repository.radenintan.ac.id Internet Source	<1%
58	www.djournals.com Internet Source	<1%
59	www.nairjc.com Internet Source	<1%
60	Ayoub Alsarhan, Mahmoud AlJamal, Osama Harfoushi, Mohammad Aljaidi et al.	<1%

"Optimizing Cyber Threat Detection in IoT: A Study of Artificial Bee Colony (ABC)-Based Hyperparameter Tuning for Machine Learning", Technologies, 2024
Publication

61	Catherine Junia, Selvan K. "Resoluteneuronet: Deep Learning-based Segmentation and Classification Covid-19 Using Chest X-ray Images", Springer Science and Business Media LLC, 2024 Publication	<1 %
62	Dewi Meta Amalya, Tri Wahyu Widyaningsih. "Implementation of Naive Bayes for Classification and Potentially MSMEs Analysis", MATEC Web of Conferences, 2018 Publication	<1 %
63	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1 %
64	Kendra Camilla Besariani, Moses Glorino Rumambo Pandin. "MORAL DEGRADATION IN THE MILLENNIAL GENERATION DUE TO MODERNIZATION", Open Science Framework, 2021 Publication	<1 %
65	Kolsoom Mehrabi, Abbas Zarifkar, Mahsa Babaei. "Compact, high-performance, and fabrication friendly two-mode division multiplexer based on a silicon bent directional coupler", Applied Optics, 2020 Publication	<1 %
66	dokumen.pub Internet Source	<1 %

67	educationaltechnologyjournal.springeropen.com Internet Source	<1%
68	edulearn.intelektual.org Internet Source	<1%
69	openpublichealthjournal.com Internet Source	<1%
70	sciengtexopen.org Internet Source	<1%
71	vtechworks.lib.vt.edu Internet Source	<1%
72	www.icicelb.org Internet Source	<1%
73	www.jait.us Internet Source	<1%
74	Elvin Elvin, Antoni Wibowo. "Forecasting water quality through machine learning and hyperparameter optimization", Indonesian Journal of Electrical Engineering and Computer Science, 2024 Publication	<1%
75	Fitriana Harahap, Ahir Yugo Nugroho Harahap, Evri Ekadiansyah, Rita Novita Sari, Robiatul Adawiyah, Charles Bronson Harahap. "Implementation of Naïve Bayes Classification Method for Predicting Purchase", 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018 Publication	<1%
76	Nindhia Hutagaol, Suharjito Suharjito. "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher	<1%

Education", Advances in Science, Technology and Engineering Systems Journal, 2019
Publication

77	"Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications", Springer Science and Business Media LLC, 2021 Publication	<1%
78	Ansar Siddique, Asiya Jan, Fiaz Majeed, Adel Ibrahim Qahmash, Noorulhasan Naveed Quadri, Mohammad Osman Abdul Wahab. "Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers", Applied Sciences, 2021 Publication	<1%
79	Caio Filipe de Lima Munguba, Gustavo de Novaes Pires Leite, Felipe Costa Farias, Alexandre Carlos Araújo da Costa et al. "Ensemble learning framework for fleet-based anomaly detection using wind turbine drivetrain components vibration data.", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
80	Jiayuan Song, Zheng Liu. "Comparison of Norm-Based Feature Selection Methods on Biological Omics Data", 2021 5th International Conference on Advances in Image Processing (ICAIP), 2021 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		