Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 3 Iss. 2 January 2026, pp: 80-88
ISSN(P): 3024-921X | ISSN(E): 3024-8043

A Comparison of Machine Learning Algorithms in Predicting Students' Academic Performance

Juanda Alra Baye[1*], Gemma Tahmid Alfaridzi[2], Hilmy Abdurrahim[3], Abid Aziz Adinda[4], Muhammad Rakha Athallah[5], Muhammad Zahid Ramadhan[6]

[1,2]Departemen of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Chemical Engineering, Faculty of Applied Science, The University of British Columbia, Canada
[4]Bachelor of Business Administration (Hons.), Faculty of Economics and Management,
Universitas Putra Malaysia, Malaysia
[5]Economics and Islamic Finance, Faculty of Economics and Islamic Studies, Yarmouk University, Jordan
[6]Syari'ah and Law, Faculty of Sharia and Law, Al-Azhar University, Egypt

E-Mail: [1]halojuandabaye@gmail.com, [2]gemmatahmid18@gmail.com, [3]hilmyabdur@gmail.com,
[4]abidaziz@gmail.com, [5]mrakhaathallah@gmail.com, [6]zahidramadhan@gmail.com

Received Dec 27th 2024; Revised Oct 15th 2025; Accepted Dec 07th 2025; Available Online Jan 31th 2026
Corresponding Author: Juanda Alra Baye

Abstract

Predicting students' academic performance enables early interventions and data-driven planning in education. We compare five machine-learning algorithms Decision Tree, K-Nearest Neighbor, Naive Bayes, Random Forest, and Support Vector Machine on a publicly available dataset of 1,001 students, evaluated with Accuracy, Precision, Recall, and F1-Score. The Decision Tree achieved the highest performance, with perfect scores on this dataset, while SVM (≈82% F1) and Random Forest (≈81% F1) were competitive. These results suggest that simple, interpretable models can be highly effective when features are clean and predictive; however, the Decision Tree's perfection also indicates potential overfitting and warrants further validation on larger, more diverse samples. The study underscores how model choice should reflect dataset characteristics and practical deployment goals in educational settings, informing early-warning systems and targeted support programs.

Keywords: Academic Performance, Decision Tree, Machine Learning, Random Forest, Support Vector Machine

1. INTRODUCTION

In the increasingly digitized era, education has also undergone major transformations through the use of advanced information technology. One of the most critical indicators in education is academic performance, which refers to students' measurable learning outcomes as reflected in test scores, grades, and other academic metrics. Understanding and predicting academic performance is essential because it not only reflects students' learning progress but also informs educational policies, resource allocation, and early interventions for students at risk. With the growing availability of digital academic records and demographic data, educational institutions now face the challenge of effectively analyzing this vast information to improve learning strategies. Traditional statistical methods often fall short in capturing complex and non-linear patterns within data. Therefore, there is an urgent need to adopt more robust and data-driven approaches that can handle large, diverse datasets and uncover complex patterns within them.

Machine learning (ML) has emerged as a powerful approach for predictive modeling, particularly due to its ability to learn patterns from large datasets. The application of ML in education allows for the identification of key predictors of student outcomes, enabling more targeted and effective interventions. This study employs five machine learning algorithms Decision Tree, K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), Random Forest, and Support Vector Machine (SVM) to predict academic performance based on student demographics and test scores. These algorithms were chosen based on their diverse characteristics and proven effectiveness in prior research. The Decision Tree algorithm is known for its clear interpretability and rule-based decision structure. KNN is widely used for its simplicity and effectiveness in small to medium-sized datasets. The Naive Bayes algorithm is favored for its speed and suitability for

categorical data. Random Forest enhances prediction accuracy through ensemble learning and reduces the risk of overfitting, while SVM is robust in handling non-linear and high-dimensional data.

The collection and analysis of students' academic data have become increasingly important practices to support evidence-based decision-making within educational systems. Data related to students' academic performance, including test scores, demographic backgrounds, and socioeconomic information, can now be accessed easily through various digital education platforms. With the growing volume of data, the main challenge is how to analyze and leverage this vast amount of data to design more effective and personalized learning strategies [1][2]. As technology and learning methods evolve, students' academic success is increasingly influenced by complex factors, including psychological, social, and economic factors [3].

Predicting students' academic performance has become a particularly intriguing topic in educational research. Various statistical approaches and machine learning (ML) algorithms have been employed to analyze academic data to predict students' success or struggles in learning. This study focuses on the use of five machine learning algorithms to predict students' academic performance, emphasizing the importance of understanding the factors influencing academic achievement in a more precise and measurable way. Most current approaches rely on traditional statistical techniques that are often limited in capturing more complex relationships between variables [4]. Therefore, machine learning algorithms have been widely adopted to uncover deeper patterns from large and heterogeneous academic data, providing more accurate predictions of students' potential and learning difficulties.

Among the various machine learning algorithms used, Decision Tree, K-Nearest Neighbor (KNN), Naive Bayes, Random Forest, and Support Vector Machine (SVM) have proven effective in predicting students's academic performance [5][6]. Each algorithm has its strengths and weaknesses in handling different types of data, and combining these methods often yields more optimal results [7]. For instance, the Decision Tree is known. for its ability to provide interpretable insights, allowing educators to understand the key factors influencing students' academic performance [8]. Meanwhile, K-Nearest Neighbor (KNN) and Naive Bayes show potential for producing quick and efficient predictions in various scenarios, although they may struggle when dealing with large or noisy data [9]. On the other hand, Random Forest and SVM perform better in handling complex and non-linear relationships, reducing the overfitting that can impair model accuracy [10][11].

Several studies have discussed the effectiveness of each of these algorithms, but few have directly compared all five algorithms in the context of predicting students' academic performance using a comprehensive and systematic approach. This study aims to introduce a more robust methodology for predicting academic performance, taking into account various factors that may influence outcomes, including psychological, social, and economic aspects of students. This aligns with the findings of Anderson et al. (2023), which show that a multi-algorithm approach can provide deeper insights and improve the reliability of predictions [12].

By combining these five algorithms in the analysis of academic data, this study aims to provide a more complete picture of how each algorithm can affect students' academic performance predictions. Additionally, the study will evaluate the efficiency, accuracy, and interpretability of the results using established evaluation metrics such as accuracy, precision, recall, and F1 score. The expected outcomes of this research can contribute significantly to the field of education, helping educational institutions design data-driven and targeted intervention programs. This research not only contributes to the development of educational data analysis theory but also offers practical solutions that can be applied to educational policies and teaching practices worldwide. With more advanced, data-driven approaches, this study aims to enhance the quality of learning and support a more inclusive and efficient educational system in the future.

2. MATERIAL AND METHOD

2.1. Stage of Research

This research methodology is designed to compare the performance of classification algorithms (C4.5, K-Nearest Neighbor, Naive Bayes, Random Forest, and Support Vector Machine) in predicting students' academic performance [13]. As seen in Figure 1.

The methodology used in this study consists of several sequential stages, as illustrated in Figure 1. The first stage is data collection, which involves obtaining a dataset from the Kaggle platform. The second stage is data preprocessing, which includes data cleaning, handling missing values, and transforming categorical data into a numerical format. This is followed by feature selection to identify the most relevant variables. The third stage is the implementation of machine learning algorithms using the Python programming language via Google Colab. Each model is trained and tested using two different data split ratios, 70:30 and 80:20, based on the holdout validation technique. The final stage is performance evaluation using classification metrics such as accuracy, precision, recall, and F1-score. The dataset used in this study is sourced from Kaggle and can be accessed via the following URL: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. It contains 1001 records of student performance data and consists of seven attributes: gender, race/ethnicity, parental level of education,

lunch type, test preparation course, and three academic scores (math score, reading score, and writing score). In this research, the main features used are: gender, race, parental education level, and test preparation course. The academic scores (math, reading, and writing) are used to determine the target class, which classifies student performance into specific categories such as high, medium, or low achievers (depending on the scoring thresholds applied in preprocessing). This classification approach allows each machine learning algorithm to predict which group a student is likely to belong to based on their attributes. Each algorithm was selected due to its relevance and past success in educational prediction tasks. The Decision Tree (C4.5) model creates a flowchart-like structure for decision-making based on entropy and Gini index. The K-Nearest Neighbor (KNN) algorithm classifies new samples based on the majority class among the k-nearest data points using Euclidean distance. The Naive Bayes classifier applies Bayes' theorem under the assumption of feature independence. Random Forest combines multiple decision trees using bagging and majority voting, which improves prediction accuracy and reduces overfitting. The Support Vector Machine (SVM) separates classes using the optimal hyperplane, especially useful for high-dimensional data and non-linear patterns. By applying the same dataset to all five algorithms, and evaluating them with identical metrics and conditions, this study ensures a fair and balanced comparison. The results will highlight not only which algorithm performs best overall but also how they behave differently under varying data proportions. The goal is to provide practical insights for researchers and educators who aim to apply predictive analytics to academic data.

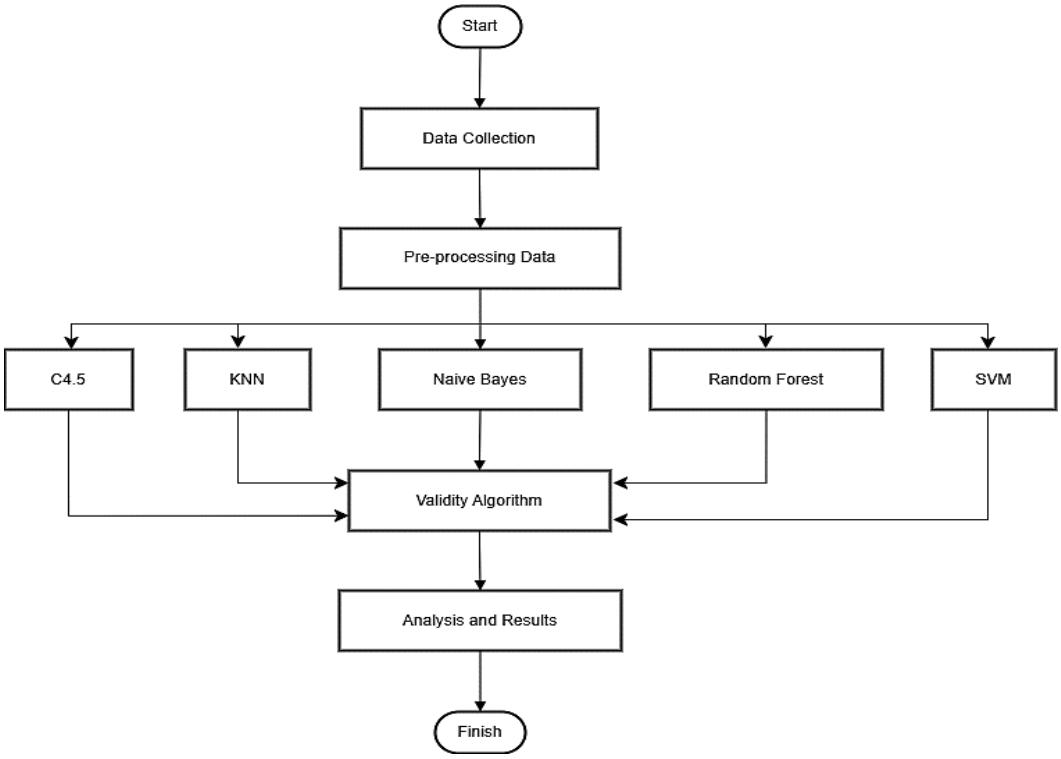

Figure 1. Research Methodology

2.2. Data Collecting

The dataset used was obtained from the Kaggle platform, containing data about students with various attributes related to their academic performance [14] . This dataset consists of 1001 records and includes 7 attributes: gender, race, parental education level, test preparation course, math scores, reading scores, and writing scores[15].

2.3. Decision Tree

The C4.5 decision tree algorithm is a machine learning model used for classification and regression, featuring a hierarchical tree structure to make predictions [16][17]. This model works by splitting the dataset into subsets based on specific features [18], resulting in decision paths that lead to target values:

1. Entropy (E), which measures the level of uncertainty in the data, is expressed by the formula 1.

$$E(S) = -\sum_{i=1}^{k} p_i \log^2(p_i) \quad (1)$$

Where p_i is the proportion of elements from class i in S [14].

2. The Gini Index, which is used to measure impurity in the data, is given by the formula 2.

$$Gini(S) = 1 - \sum_{i=1}^{k} \quad {p_i}^2 \tag{2}$$

A lower Gini Index indicates a cleaner separation at the tree node [19].

2.4. K- Nearest Neighbor

The K-nearest neighbor (KNN) algorithm is a non-parametric learning technique widely used for classification and regression tasks [20][21]. KNN operates under the assumption that data points close to each other in the feature space tend to share similar characteristics, classifying new data points based on the majority label of the 'k' nearest neighbors [22]. This algorithm often uses the Euclidean distance to measure the closeness between data points [23], which can be calculated using the following formula 3.

$$D(X,Y) = \sqrt{\left\{\sum_{\{i=1\}}^{n} \quad (x_i - y_i)^2\right\}} \tag{3}$$

Here, $D(X,Y)$ represents the Euclidean distance between two points X and Y, while x_i and y_i are the coordinates of each dimension in the feature space [2].

2.5. Naïve Bayes Classifier

The Naive Bayes Classifier (NBC) is a probability-based classification algorithm that applies Bayes' Theorem under the assumption that all features are independent within a given class [24][25]. This algorithm calculates conditional probabilities for each class based on the available features, enabling fast and effective classification, particularly for categorical data [26], as shown in equation 4.

$$P(X) = \frac{P(C)\cdot P(C)}{P(X)} \tag{4}$$

Where $P(X)$ is the probability of class C given feature X, $P(X)$ is the probability of a feature X given class C,and $P(C)$ is the prior probability of class C [27].

2.6. Random Forest

The ensemble-based machine learning method Random Forest (RF) uses multiple decision trees to increase prediction accuracy [28]. RF is used in this study to forecast kids' academic achievement. The bootstrap method is used to construct each tree from distinct subsets of data, and the final result is generated by averaging for regression or majority voting for classification [29][30]. Applications like geographical categorization and agronomic predictions benefit greatly from RF's ability to handle non-linear data and a variety of variables [31]. The impurity reduction at each feature is measured to determine feature importance, or Gini Importance.. The feature importance $I(f)$ for feature i in tree t is formulated 5.

$$I(f) = \frac{1}{T}\sum_{\{t=1\}}^{T} \quad \Delta I(t,f) \tag{5}$$

Where T is the total number of trees in the forest and $\Delta I(t,f)$ is the impurity reduction in tree t caused by feature f [12].

2.7. Support Vector Machine

Support Vector Machine (SVM) is a statistical-based machine learning algorithm that separates data using an optimal hyperplane, effective for both classification and regression tasks [5][32]. SVM maximizes the margin between two classes, making it particularly useful for datasets that are not linearly separable [33]. The Optimal Hyperplane formula for predicting data class is equation 6 [35].

$$Prediction = W^T X + b \tag{6}$$

Where W is the weight vector, X is the feature vector of the sample, and b is the bias.

3. RESULTS AND ANALYSIS

The results of this study demonstrate the performance of five machine learning algorithms in predicting students' academic performance using demographic information and academic test scores. The dataset, consisting of 1001 records, was preprocessed and split into training and testing data using two

proportions: 70:30 and 80:20. Each model was evaluated using four performance metrics: accuracy, precision, recall, and F1-score. The algorithms' outcomes are summarized and analyzed in this section.

3.1 Collecting and Preprocessing Data

The results of this study demonstrate the performance of five machine learning algorithms in predicting students' academic performance based on demographic data and test scores. The dataset used contains 1001 records with 7 attributes, including gender, race, parental education level, and math, reading, and writing test scores. After undergoing data preprocessing, including data cleaning and feature selection to choose the most relevant variables, the dataset was divided into two parts using the Holdout Split method with a 70:30 ratio for training and testing data. The five algorithms (C4.5 Decision Tree, K-Nearest Neighbor, Naive Bayes, Random Forest, and Support Vector Machine) were tested based on accuracy, precision, recall, and F1-score to evaluate their effectiveness in predicting students' academic performance.

3.2 Process Validaty Algorithm

The prediction implementation in this study was carried out using five models: C4.5 Decision Tree, K-Nearest Neighbor, Naive Bayes, Random Forest, and Support Vector Machine. These models processed 1001 student data that had undergone preprocessing stages.

3.2.1 Decesion Tree

The Decision Tree (DT) algorithm achieved the highest performance across both data splits, with precision, recall, F1-score, and accuracy all reaching 100%. These perfect scores suggest that the model completely separated the classes in the dataset, with no misclassifications. This result indicates the Decision Tree's strong ability to interpret the dataset, although it may also suggest the possibility of overfitting, particularly given the relatively small dataset size. The validation results of the DT algorithm are in Table 1.

Table 1. Performa DT

Split Data	Precision	Recall	F1-Score	Accuracy
70:30	100%	100%	100%	100%
80:20	100%	100%	100%	100%

3.2.2 K-Nearest Neighbors (K-NN)

The K-Nearest Neighbor (KNN) algorithm showed moderate performance. On the 70:30 split, it achieved 70.51% precision, 77.66% recall, and 73.91% F1-score with 64.00% accuracy. When trained on 80% of the data, performance improved slightly, with an F1-score of 77.70% and an accuracy of 69.00%. These results demonstrate that KNN benefits from more training data but still lacks precision, likely due to its sensitivity to noise and feature scale. The validation results of the K-NN algorithm are in Table 2.

Table 2. Performa K-NN

Split Data	Precision	Recall	F1-Score	Accuracy
70:30	70,51%	77,66%	73,91%	64,00%
80:20	72,00%	84,38%	77,70%	69,00%

3.2.3 Naïve Bayes Classifier

The Naive Bayes Classifier (NBC) displayed relatively stable but moderate results, with F1-scores around 73% for both data splits and an accuracy of approximately 65%. This consistency reflects the algorithm's robustness, especially with categorical data. However, the assumption of feature independence likely limited its ability to capture more complex relationships among attributes. The validation results of the Naive Bayes Classifier algorithm are in Table 3.

Table 3. Performa NBC

Split Data	Precision	Recall	F1-Score	Accuracy
70:30	74,23%	73,10%	73,66%	65,67%
80:20	72,87%	73,44%	73,15%	65,50%

3.2.4 Random Forest

The Random Forest (RF) algorithm yielded strong results, particularly on the 80:20 split, achieving an F1-score of 81.43% and 74.00% accuracy. This performance can be attributed to the model's ensemble structure, which combines multiple decision trees to enhance generalization and reduce overfitting. It performed slightly better than SVM on the 80:20 split but was outperformed by SVM on the 70:30 split. The validation results of the Random Forest algorithm are in Table 4.

Table 4. Performa RF

Split Data	Precision	Recall	F1-Score	Accuracy
70:30	74,68%	88,32%	80,93%	72,67%
80:20	75,00%	89,06%	81,43%	74,00%

3.2.5 Support Vector Machine

The Support Vector Machine (SVM) also showed competitive results, with an F1-score of 82.08% and accuracy of 74.67% on the 70:30 split. The model maintained consistent performance across both splits, indicating its robustness in handling non-linear data patterns. SVM's ability to maximize the margin between classes proved effective, particularly with non-linear relationships represented through one-hot encoded categorical features. The validation results of the Support Vector Machine (SVM) algorithm are in Table 5.

Table 5. Performa SVM

Split Data	Precision	Recall	F1-Score	Accuracy
70:30	76,65%	88,32%	82,08%	74,67%
80:20	74,84%	90,62%	81,98%	74,50%

3.3 Comparison Results of Algorithms

After evaluating and testing the performance of the algorithms on each data split, the next step is to compare them to determine the algorithm with the best performance. The results of this comparison are shown in Figure 2.

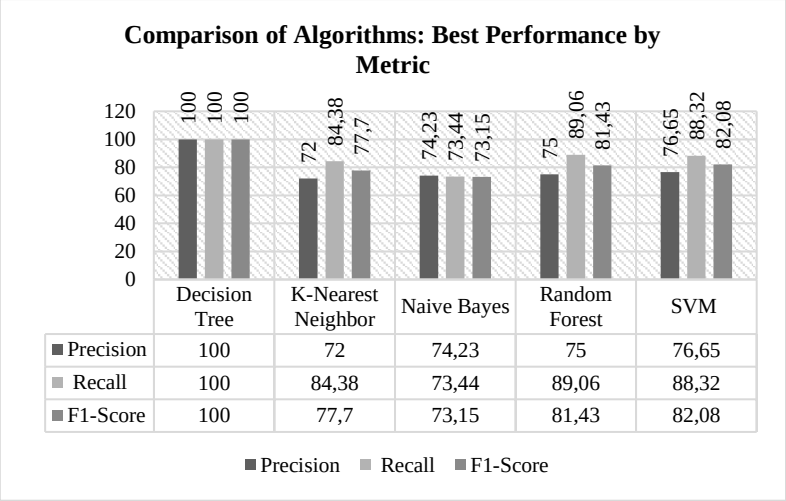

Figure 2. Comparison of Algorithm

The K-Nearest Neighbor (K-NN) algorithm showed the best performance on the 80:20 split, achieving a Precision of 72%, Recall of 84%, and F1-Score of 78%. Naive Bayes Classifier (NBC) also provided strong results on the 70:30 split, with a Precision of 74%, Recall of 73%, and F1-Score of 73%. However, the Decision Tree (DT) algorithm achieved perfect results with 100% for all evaluation metrics (Precision, Recall, and F1-Score) across both data splits. To determine the top-performing algorithm among these three, further comparison is carried out, as shown in Figure 3.

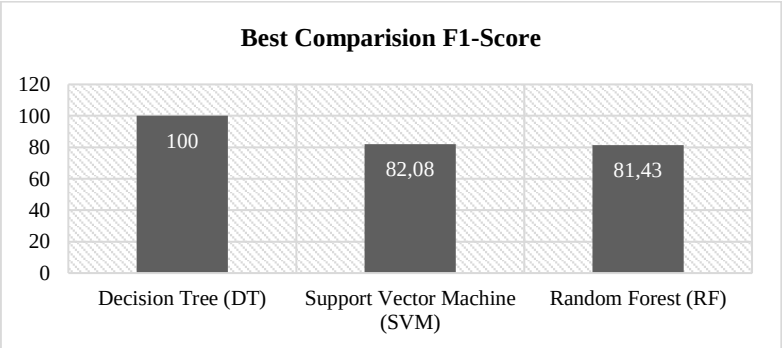

Figure 3. Best Comparision F1-Score

The Decision Tree (DT) method performs the best, with an F1-Score of 100%, according to the F1-Score values for the three algorithms shown in Figure 3. With an F1-Score of 82.08%, the Support Vector Machine (SVM) comes in second, while the Random Forest (RF) comes in third with an F1-Score of 81.43%. Therefore, based on its F1-Score performance, the Decision Tree (DT) method is determined to be the most optimal of the three in this study.

3.4 Discussion

To contextualize these results, prior studies such as those by Zhang et al. (2023) and Rajakumaran et al. (2024) have found similar trends, where Decision Tree and SVM consistently outperform simpler models like KNN or NBC in educational datasets. Our findings reinforce these conclusions, showing that while all five algorithms can be used for academic performance prediction, Decision Tree remains the most effective in this specific dataset. However, our perfect scores for Decision Tree also raise questions about generalizability, and further testing on larger or more diverse datasets is recommended. In comparison with related studies, the uniqueness of this research lies in the direct side-by-side evaluation of five algorithms under consistent preprocessing and evaluation standards. While other research often focuses on a single algorithm or uses varied datasets, this study controls the experiment to draw clear performance distinctions. One limitation is the reliance on a single dataset with limited diversity, which may restrict the model's general applicability. Moreover, the dataset does not include temporal or behavioral attributes (such as attendance, participation, or motivation), which could enhance prediction depth.

The Decision Tree's perfect accuracy likely reflects deterministic splits induced by low-noise, predominantly categorical features (e.g., gender, parental education, and test-preparation) combined with label thresholds derived from exam scores. Such feature–label structure can yield pure nodes when class boundaries align with simple rules. Nevertheless, achieving 100% on hold-out splits in a 1,001-record dataset remains uncommon; hence, we treat this as a red flag for potential overfitting and assess it alongside additional validation recommendations.

Perfect metrics suggest that the model may memorize dataset-specific patterns rather than learn generalizable rules. We therefore recommend k-fold cross-validation (e.g., stratified 5×2 CV), post-pruning, and evaluation on external datasets to ensure robustness. For tree-based models, setting limits on maximum depth and minimum samples per leaf can mitigate overfitting risks. By "more complex data," we refer to settings with non-linear decision boundaries and higher-order interactions among predictors. Kernel SVM (e.g., RBF) maps inputs into a higher-dimensional space, enabling linear separation there while preserving a large margin. This typically benefits datasets where categorical and continuous attributes interact in non-additive ways. In practice, careful feature scaling and hyperparameter tuning (C, γ) are crucial for robust performance. These results highlight the importance of aligning model choice with dataset characteristics and overfitting risks [2], [10], [12].

From a practical standpoint, the findings of this study can help educational institutions or data-driven learning platforms to implement early warning systems. For instance, identifying students with predicted low performance can prompt interventions such as remedial programs, mentoring, or counseling. The predictive accuracy demonstrated by the Decision Tree model makes it a suitable candidate for real-time applications in academic dashboards or intelligent tutoring systems. In conclusion, the comparative results show that Decision Tree provides the most accurate prediction, followed closely by SVM and Random Forest. While KNN and Naive Bayes offer simplicity and speed, their lower precision and accuracy suggest they may be more suitable for preliminary analysis or in environments with constrained computational resources. Future studies are encouraged to explore algorithm optimization through hyperparameter tuning, use of ensemble models, or application on more complex and real-world academic datasets. This research contributes by providing a direct, standardized comparison among five widely used machine-learning algorithms on the same educational dataset, something rarely done in prior studies. Unlike previous works that focus on a single algorithm, this study highlights how algorithmic complexity and data characteristics jointly affect predictive accuracy. The insights help educational data scientists choose the most appropriate model depending on data type, noise level, and feature dimensionality.

4. CONCLUSION

Based on the analysis and evaluation of five machine learning algorithms—Decision Tree, K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), Random Forest, and Support Vector Machine (SVM)—this study concludes that the Decision Tree algorithm is the most effective in predicting students' academic performance. It achieved perfect scores in all evaluation metrics (precision, recall, F1-score, and accuracy) across both data splits (70:30 and 80:20). This result highlights its strong capability in capturing the structure of the dataset and making precise classifications. Among the other algorithms, Support Vector Machine and Random Forest also showed promising results, with F1-scores above 80%, demonstrating their reliability for predictive tasks involving educational data. In contrast, Naive Bayes and KNN, while computationally efficient and easy to implement, yielded lower predictive performance and may be more suitable for

exploratory analysis rather than high-stakes academic forecasting. These findings offer practical value for educational institutions aiming to adopt predictive models to identify students who may be at risk of poor academic performance. By leveraging accurate prediction models such as Decision Tree, schools and universities can implement early interventions, design personalized learning strategies, and allocate resources more effectively to improve student outcomes.

Although the Decision Tree achieved perfect metrics, this performance may not generalize beyond the current dataset. The dataset's limited diversity and the absence of behavioral or temporal variables constrain predictive scope and may introduce bias. Future research should apply cross-validation, post-pruning, and external validation using multi-institutional data. It is also recommended to include engagement and attendance features and perform hyperparameter tuning for robustness.

Despite these limitations, this study demonstrates that simple, interpretable algorithms such as Decision Tree can yield highly accurate predictions when features are clean and relevant. These findings provide practical guidance for institutions developing early-warning systems, enabling timely academic interventions based on reliable data insights. However, ensuring model generalization and transparency remains crucial before applying such models to real educational environments.. The importance of this study lies in its potential impact on educational decision-making. By demonstrating that machine-learning models particularly Decision Tree can accurately forecast academic outcomes, institutions can design early-warning systems and allocate resources more efficiently. In this sense, the research not only advances predictive modeling but also supports data-driven educational policies that promote equity and student success.

REFERENCES

[1] L. Smith and C. Lamprecht, "Identifying the limitations associated with machine learning techniques in performing accounting tasks," vol. 22, no. 2, pp. 227–253, 2024, doi: 10.1108/JFRA-05-2023-0280.

[2] S. Zhang, S. Member, and J. Li, "KNN Classification With One-Step Computation," IEEE Trans. Knowl. Data Eng., vol. 35, no. 3, pp. 2711–2723, 2023, doi: 10.1109/TKDE.2021.3119140.

[3] I. Lillo-bravo, J. Vera-medina, C. Fernandez-peruchena, and E. Perez-aparicio, "Random Forest model to predict solar water heating system performance," Renew. Energy, vol. 216, no. April, p. 119086, 2023, doi: 10.1016/j.renene.2023.119086.

[4] S. Lee, "Transformation Based Tri-Level Feature Selection Approach Using Wavelets and Swarm Computing for Prostate Cancer Classification," vol. 8, 2020.

[5] M. Rajakumaran, G. Arulselvan, S. Subashree, and R. Sindhuja, "Measurement : Sensors Crop yield prediction using multi-attribute weighted tree-based support vector machine," Meas. Sensors, vol. 31, no. May 2023, p. 101002, 2024, doi: 10.1016/j.measen.2023.101002.

[6] I. J. A. Res, "Manuscript Info Abstract ISSN : 2320-5407 Introduction : -," vol. 12, no. 01, pp. 422–438, 2024, doi: 10.21474/IJAR01/18138.

[7] I. Bagus, G. Purwania, I. N. S. Kumara, and M. Sudarma, "Application of IoT-Based System for Monitoring Energy Consumption," vol. 5, no. 2, 2020.

[8] "Par : Henri Barki , École des HEC Jon Hartwick , McGill University," no. June 2001, 2014, doi: 10.2307/3250929.

[9] S. Chowdhury, "Comparison of accuracy and reliability of random forest , support vector machine , artificial neural network and maximum likelihood method in land use / cover classification of urban setting," Environ. Challenges, vol. 14, no. October 2023, p. 100800, 2024, doi: 10.1016/j.envc.2023.100800.

[10] A. Coscia, V. Dentamaro, S. Galantucci, A. Maci, and G. Pirlo, "Journal of Information Security and Applications Automatic decision tree-based NIDPS ruleset generation for DoS / DDoS attacks," J. Inf. Secur. Appl., vol. 82, no. March, p. 103736, 2024, doi: 10.1016/j.jisa.2024.103736.

[11] Z. Zhao, Z. Luo, J. Li, C. Chen, and Y. Piao, "When Self-Supervised Learning Meets Scene Classification : Remote Sensing Scene Classification Based on a Multitask Learning Framework," pp. 1–22, 2020, doi: 10.3390/rs12203276.

[12] X. Zhang, H. Shen, T. Huang, Y. Wu, B. Guo, and Z. Liu, "Improved random forest algorithms for increasing the accuracy of forest aboveground biomass estimation using Sentinel-2 imagery," Ecol. Indic., vol. 159, no. February, p. 111752, 2024, doi: 10.1016/j.ecolind.2024.111752.

[13] S. García-ponsoda, A. Maté, and J. Trujillo, "Refining ADHD diagnosis with EEG : The impact of preprocessing and temporal segmentation on classification accuracy," Comput. Biol. Med., vol. 183, no. September, p. 109305, 2024, doi: 10.1016/j.compbiomed.2024.109305.

[14] J. Mushava and M. Murray, "Flexible loss functions for binary classification in gradient-boosted decision trees : An application to credit scoring," Expert Syst. Appl., vol. 238, no. PC, p. 121876, 2024, doi: 10.1016/j.eswa.2023.121876.

[15] N. Liu, Y. Xiang, F. Wang, and S. Cao, "Big Data Course Multidimensional Evaluation Model based on Knowledge Graph enhanced Transformer," Cogn. Robot., 2024, doi: 10.1016/j.cogr.2024.11.003.

[16] F. Farah, A. Ahmed, and R. Ça, “Results in Engineering Integrating autoencoder and decision tree models for enhanced energy consumption forecasting in microgrids : A meteorological data-driven approach in Djibouti,” vol. 24, no. September, 2024, doi: 10.1016/j.rineng.2024.103033.
[17] D. Ananda, S. Nurhidayarnis, and T. A. Afifah, “Text Classification of Translated Qur ’ anic Verses Using Supervised Learning Algorithm,” vol. 1, no. January, pp. 78–84, 2024.
[18] T. Hara and M. Sasabe, “Practicality of in-kernel / user-space packet processing empowered by lightweight neural network and decision tree ☆,” Comput. Networks, vol. 240, no. December 2023, p. 110188, 2024, doi: 10.1016/j.comnet.2024.110188.
[19] S. Devasahayam and B. Albijanic, “Predicting hydrogen production from co-gasification of biomass and plastics using tree based machine learning algorithms,” Renew. Energy, vol. 222, no. November 2023, p. 119883, 2024, doi: 10.1016/j.renene.2023.119883.
[20] S. S. Shijer, A. H. Jassim, L. A. Al-haddad, and T. T. Abbas, “e-Prime - Advances in Electrical Engineering , Electronics and Energy Evaluating electrical power yield of photovoltaic solar cells with k-Nearest neighbors : A machine learning statistical analysis approach,” e-Prime - Adv. Electr. Eng. Electron. Energy, vol. 9, no. June, p. 100674, 2024, doi: 10.1016/j.prime.2024.100674.
[21] A. Maity, P. Prakasam, and S. Bhargava, “Robust dual-tone multi-frequency tone detection using k-nearest neighbour classifier for a noisy environment,” 2020, doi: 10.1108/ACI-10-2020-0105.
[22] J. Kim, J. Choi, Y. Park, C. K. Leung, S. Member, and A. Nasridinov, “KNN-SC : Novel Spectral Clustering Algorithm Using k-Nearest Neighbors,” IEEE Access, vol. 9, pp. 152616–152627, 2021, doi: 10.1109/ACCESS.2021.3126854.
[23] Y. Peng, “LK-Index : A Learned Index for KNN Queries,” IEEE Access, vol. 12, no. August, pp. 103096–103103, 2024, doi: 10.1109/ACCESS.2024.3433524.
[24] C. K. Zarry and A. Kurniawan, “Classifications of Offline Shopping Trends and Patterns with Machine Learning Algorithms,” vol. 2, no. July, pp. 18–25, 2024.
[25] Z. Xue, J. Wei, and W. Guo, “A Real-Time Naive Bayes Classifier Accelerator on FPGA,” IEEE Access, vol. 8, pp. 40755–40766, 2020, doi: 10.1109/ACCESS.2020.2976879.
[26] C. J. Anderson et al., “A novel naïve Bayes approach to identifying grooming behaviors in the force-plate actometric platform,” J. Neurosci. Methods, vol. 403, no. July 2023, p. 110026, 2024, doi: 10.1016/j.jneumeth.2023.110026.
[27] Q. A. Al-haija and A. A. Alsulami, “Fast anomalous traffic detection system for secure vehicular communications,” vol. 5, no. x, 2024, doi: 10.23919/ICN.2024.0021.
[28] K. Sumwiza, C. Twizere, G. Rushingabigwi, and P. Bakunzibake, “Informatics in Medicine Unlocked Enhanced cardiovascular disease prediction model using random forest algorithm,” Informatics Med. Unlocked, vol. 41, no. August, p. 101316, 2023, doi: 10.1016/j.imu.2023.101316.
[29] K. A. Mahasiswa, R. Rachmatika, and A. Bisri, “Perbandingan Model Klasifikasi untuk Evaluasi,” vol. 6, no. 3, pp. 417–422, 2020.
[30] C. P. Trisya, N. W. Azani, and L. M. Sari, “Performance Comparison of ARIMA , LSTM and SVM Models for Electric Energy Consumption Analysis,” vol. 1, no. January, pp. 85–94, 2024.
[31] E. Asamoah, G. B. M. Heuvelink, and I. Chairi, “Heliyon Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana,” vol. 10, no. July, 2024.
[32] A. Shell, “Optimization of hydrochar production from almond shells using response surface methodology, artificial neural network, support vector machine and XGBoost,” Desalin. Water Treat., vol. 317, no. February, p. 100154, 2024, doi: 10.1016/j.dwt.2024.100154.
[33] Y. Zhu, C. Gu, and M. A. Diaconeasa, “A missing data processing method for dam deformation monitoring data using spatiotemporal clustering and support vector machine model,” Water Sci. Eng., vol. 17, no. 4, pp. 417–424, 2024, doi: 10.1016/j.wse.2024.08.003.
[34] Kaggle, "Students Performance in Exams Dataset", https://www.kaggle.com/datasets/spscientist/students-performance-in-exams, accessed on May 2025.
[35] V. Wulandari, Mustakim, R. Novita and N. E. Rozanda, "Implementation of Machine Learning Algorithm for Stroke Risk Classification by Applying Sequential Forward Selection," 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI), Jakarta, Indonesia, 2025, pp. 696-701, doi: 10.1109/ICoCSETI63724.2025.11020494.