Original Article

Check for updates

Available online at
www.heca-analitika.com/jeml

Journal of Educational Management and Learning

Vol. 3, No. 2, 2025

The Role of Study Habits, Parental Involvement, and School Environment in Predicting Student Achievement: A Machine Learning Perspective

Teuku Rizky Noviandy [1], Maria Paristiowati [2], Illyas Md Isa [3] and Rinaldi Idroes [4,*]

[1] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)
[2] Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta 13220, Indonesia; maria.paristiowati@unj.ac.id (M.P.)
[3] Department of Chemistry, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Perak 35900, Malaysia; illyas@fsmt.upsi.edu.my (I.M.I.)
[4] School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

* Correspondence: rinaldi.idroes@usk.ac.id

Article History

Received 22 August 2025
Revised 26 October 2025
Accepted 3 November 2025
Available Online 15 November 2025

Keywords:
Machine learning
Student achievement
Predictive analytics
Educational data mining

Abstract

This study explores the application of machine learning techniques to predict student achievement based on study habits, parental involvement, and school environment. Using a dataset from Kaggle comprising academic, behavioral, and contextual variables, four machine learning algorithms, namely K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest, were implemented and evaluated. Model performance was evaluated using accuracy, precision, recall, F1-score, ROC curve, and Precision–Recall curves. Results show that all models effectively classified students into low- and high-achievement categories, with SVM achieving the highest accuracy (94.02%) and the strongest overall performance. The findings highlight the potential of machine learning-driven predictive analytics in educational settings, enabling early identification of at-risk students and supporting evidence-based interventions. By integrating diverse factors influencing academic performance, this study demonstrates how data-driven approaches can enhance educational management, inform policy, and promote equitable learning outcomes.

1. Introduction

Student achievement has long been recognized as a central indicator of educational effectiveness and societal progress [1]. Across various educational systems, improving student learning outcomes remains a primary goal for educators, policymakers, and researchers. Academic success is influenced by a wide range of factors that extend beyond the classroom, including students' study behaviors, family background, and school environment [2]. Understanding how these diverse elements interact to shape learning performance is crucial for developing evidence-based strategies that improve both teaching quality and student support. In an era where data-driven decision-making is increasingly integrated into education, identifying the predictors of academic achievement has become a key focus of modern educational research.

Previous studies have examined numerous factors that influence student achievement, including socioeconomic status, parental education, access to learning resources,

and intrinsic motivation [3–6]. Similarly, institutional factors like teacher quality, school infrastructure, and peer influence have also been linked to variations in academic performance. However, traditional statistical methods often struggle to capture the complex, nonlinear relationships among these multidimensional variables [7]. With the expansion of educational data sources, there is now a growing opportunity to apply computational approaches that can analyze these intricate patterns more effectively.

Despite the availability of extensive educational data, many institutions still rely on conventional assessment tools and descriptive analyses that provide limited predictive power [8]. These methods often fail to account for the complex interactions between behavioral, social, and environmental factors that influence student outcomes [9]. As a result, schools and educators may lack the analytical capacity to identify at-risk students early or to design targeted interventions tailored to individual needs [10]. This analytical gap underscores the need to adopt more sophisticated predictive models that can process complex data structures and generate actionable insights to enhance student achievement.

Machine learning, a subset of artificial intelligence, offers a robust framework for analyzing large and heterogeneous datasets to uncover hidden patterns and relationships [11]. Machine learning algorithms are capable of learning from data without explicit programming, enabling them to model nonlinear dependencies and make accurate predictions [12–14]. In the educational context, machine learning techniques have been successfully applied to predict academic performance, detect learning difficulties, personalize instruction, and support data-driven decision-making [15–17]. By using these methods, researchers and educators can move beyond simple analyses to develop data-driven strategies that improve learning outcomes. Thus, machine learning provides a practical approach to overcoming the limitations of traditional methods in understanding and predicting student achievement.

This study aims to investigate the extent to which study habits, parental involvement, and school environment contribute to predicting student achievement using machine learning techniques. Specifically, it aims to develop and evaluate predictive models that categorize students into achievement levels based on a combination of academic, behavioral, and contextual factors. This study compares four machine learning algorithms, K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest to determine which model most effectively predicts student achievement. The goal is to provide empirical insights that can help educators, administrators, and policymakers make informed decisions supported by predictive analytics.

The primary contribution of this research lies in integrating machine learning methods with educational data to uncover the multifactorial determinants of student achievement. Unlike traditional analyses that often focus on isolated variables, this study adopts a holistic and data-driven perspective, capturing the interplay among study habits, family dynamics, and school conditions. Moreover, the use of multiple algorithms allows for a comparative understanding of model performance and robustness, highlighting which techniques are best suited for educational prediction tasks. The findings not only demonstrate the applicability of machine learning in educational management but also offer actionable insights for developing early intervention systems and data-informed policies to enhance academic success.

The remainder of this paper is organized as follows. Section 2 describes the materials and methods, including details of the dataset, machine learning algorithms, and evaluation procedures used in the analysis. Section 3 presents the results and discussion, summarizing model performance, interpreting key findings, outlining the educational implications, discussing study limitations, and suggesting directions for future research. Finally, Section 4 concludes the paper with a summary of the main insights and their significance for educational management and policy development.

2. Materials and Methods

2.1. Dataset

The dataset used in this study is the "Student Performance Factors" dataset, obtained from Kaggle [18]. This dataset is synthetic, meaning it was generated for educational and analytical purposes rather than collected from real-world institutions. Despite its synthetic nature, the dataset was designed to mimic realistic patterns of student performance by integrating various academic, personal, and environmental factors that influence achievement. It contains both quantitative and qualitative variables, making it well-suited for machine learning–based prediction and analysis of factors affecting student outcomes. Moreover, the use of a synthetic dataset ensures data privacy and provides a controlled environment for methodological validation.

The dataset includes demographic, behavioral, and contextual attributes covering aspects of study habits, parental background, school environment, and lifestyle factors. Table 1 summarizes the attributes and their descriptions.

Table 1. Attributes in the dataset.

Attribute	Description
Hours_Studied	Number of hours spent studying per week.
Attendance	Percentage of classes attended.
Parental_Involvement	Level of parental involvement (Low, Medium, High).
Access_to_Resources	Availability of educational resources (Low, Medium, High).
Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
Sleep_Hours	Average hours of sleep per night.
Previous_Scores	Student's previous exam scores.
Motivation_Level	Student's motivation (Low, Medium, High).
Internet_Access	Availability of internet access (Yes, No).
Tutoring_Sessions	Number of tutoring sessions per month.
Family_Income	Family income level (Low, Medium, High).
Teacher_Quality	Teacher quality (Low, Medium, High).
School_Type	Type of school (Public, Private).
Peer_Influence	Peer influence on academics (Positive, Neutral, Negative).
Physical_Activity	Hours of physical activity per week.
Learning_Disabilities	Presence of learning disabilities (Yes, No).
Parental_Education_Level	Highest parental education (High School, College, Postgraduate).
Distance_from_Home	Distance to school (Near, Moderate, Far).
Gender	Gender of the student (Male, Female).
Exam_Score	Final exam score.

For classification purposes, the exam score variable was transformed into two achievement categories: low and high, based on predefined thresholds. Scores below 65 were classified as low, while scores equal to or above 65 were classified as high. The resulting distribution indicates that the majority of students fall into the low category, with 4,982 students, followed by 1,625 in the high category.

2.2. Data Preprocessing

Prior to model development, the dataset underwent a comprehensive preprocessing pipeline to ensure data consistency and suitability for machine learning analysis. The preprocessing steps included data cleaning, categorical variable encoding, feature scaling, and partitioning into training and testing subsets [19].

Initially, the dataset was examined for missing or inconsistent values; no null entries were detected, confirming data completeness. Since the dataset contained both categorical and numerical attributes, appropriate transformations were applied to convert all features into a machine-readable numerical format. Categorical variables were label-encoded, assigning unique integer values to each category while preserving their distinct identities.

To normalize the range of numerical features and mitigate the effects of scale disparities, continuous variables were standardized so that each had a mean of zero and a standard deviation of one [20]. This step enhanced the convergence and performance of machine learning algorithms. Finally, the preprocessed dataset was divided into training and testing sets using an 80:20 split ratio, with 80% of the data allocated for model training and 20% reserved for evaluation [21].

2.3. Machine Learning Methods

In this study, four widely used machine learning algorithms were applied to classify student achievement levels: K-KNN, Naïve Bayes, SVM, and Random Forest. All models were implemented using the scikit-learn library with default hyperparameter settings. A brief overview of each method is presented below.

2.3.1. K-Nearest Neighbors (KNN)

KNN is a simple, instance-based learning algorithm that classifies a new observation by considering the majority class of its k nearest neighbors in the feature space [22]. It does not assume any underlying data distribution and works effectively when classes are well-separated. However, its performance can be affected by noisy features and the choice of distance metric.

2.3.2. Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, with the simplifying assumption that features are conditionally independent given the target class [23]. Despite this assumption often being unrealistic, it performs well in many real-world tasks, especially with high-dimensional data. It is computationally efficient and provides quick predictions.

2.3.3. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that aims to find the optimal hyperplane that separates classes with the maximum margin [24]. It is effective for both linear and

non-linear classification through the use of kernel functions. SVM is known for its robustness in handling high-dimensional datasets, although it may require careful tuning of parameters such as the kernel type and regularization.

2.3.4. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their predictions through majority voting [25]. Aggregating results from several trees reduces overfitting and improves generalization. Random Forest also provides insights into feature importance, making it a valuable tool for interpretability.

2.4. Performance Evaluation

The performance of the machine learning models was evaluated using four standard metrics for binary classification: accuracy, precision, recall, and F1-score [26–28]. Accuracy measures the overall proportion of correctly classified instances, providing a general indication of model performance. Precision reflects the proportion of correctly predicted positive cases among all predicted positives, while recall measures the proportion of actual positive cases correctly identified by the model. The F1-score, which is the harmonic mean of precision and recall, offers a balanced assessment of both metrics, especially when class distributions are uneven.

In addition to these metrics, two graphical methods were used to assess model performance further: the Receiver Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve. The ROC curve illustrates the trade-off between the true positive rate and false positive rate across different threshold values [29]. In contrast, the area under the curve (AUC) provides a single value summarizing the model's discriminative ability. The PR curve, on the other hand, highlights the relationship between precision and recall, offering more informative insights in the presence of class imbalance [30]. Together, these evaluation methods provide a comprehensive understanding of each model's predictive capability and robustness in classifying students into low and high-achievement categories.

3. Results and Discussion

The performance of the four machine learning models is summarized in Table 2, which presents the accuracy, precision, recall, and F1-score for each classifier.

KNN achieved an accuracy of 84.86%, with precision, recall, and F1-score values also in the mid-80s. While KNN demonstrated reasonable effectiveness in classifying student achievement levels, its reliance on distance-based similarity likely made it more sensitive to overlapping classes and noise within the dataset. This sensitivity may have led to misclassifications, especially in cases where students with similar study habits or environmental conditions had differing outcomes. Despite these limitations, KNN provided a useful baseline for comparison with more complex models.

Naïve Bayes, SVM, and Random Forest each demonstrated progressively improved performance. Naïve Bayes attained an accuracy of 89.64%, effectively capturing general patterns in the data despite its simplifying independence assumption. SVM achieved the highest overall accuracy of 94.02%, leveraging its capacity to identify an optimal separating hyperplane in the multidimensional feature space. Random Forest followed closely with an accuracy of 91.91%, combining multiple decision trees to reduce overfitting and enhance prediction stability. Overall, while all four models performed well, SVM emerged as the most effective approach for predicting student achievement, showcasing its robustness and adaptability to complex educational datasets.

Figure 1 presents the confusion matrices for the four classifiers: (a) KNN, (b) Naïve Bayes, (c) SVM, and (d) Random Forest. These visualizations offer valuable insights into how each model classified students into low and high achievement categories. The KNN model correctly identified most low-achieving students (925), but it misclassified 154 high achievers as low. This suggests that KNN struggled to differentiate between higher-performing students, likely due to overlapping feature values among the groups. Its reliance on distance-based similarity made it more sensitive to class overlap and feature noise, resulting in a higher rate of misclassification for high achievers.

The Naïve Bayes model improved the identification of high achievers, correctly classifying 251 of them while misclassifying 97 as low. Although this marked an improvement over KNN, Naïve Bayes still tended to underpredict high performance, likely because its assumption of feature independence is unrealistic in complex educational datasets with correlated factors. The SVM classifier delivered the best results, correctly classifying 952 low and 291 high achievers, with only 57 misclassified, demonstrating the strength of its margin maximization approach in handling multidimensional data. The Random Forest model also performed strongly, correctly predicting 955 low achievers and 260 high achievers, although 88 high achievers were misclassified as low achievers. While slightly less precise than SVM in capturing the high group, Random Forest's ensemble

Table 2. Performance comparison of machine learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN	84.86	84.18	84.64	83.67
Naïve Bayes	89.64	89.45	89.64	89.32
SVM	94.02	93.99	94.02	93.92
Random Forest	91.91	91.99	91.91	91.60

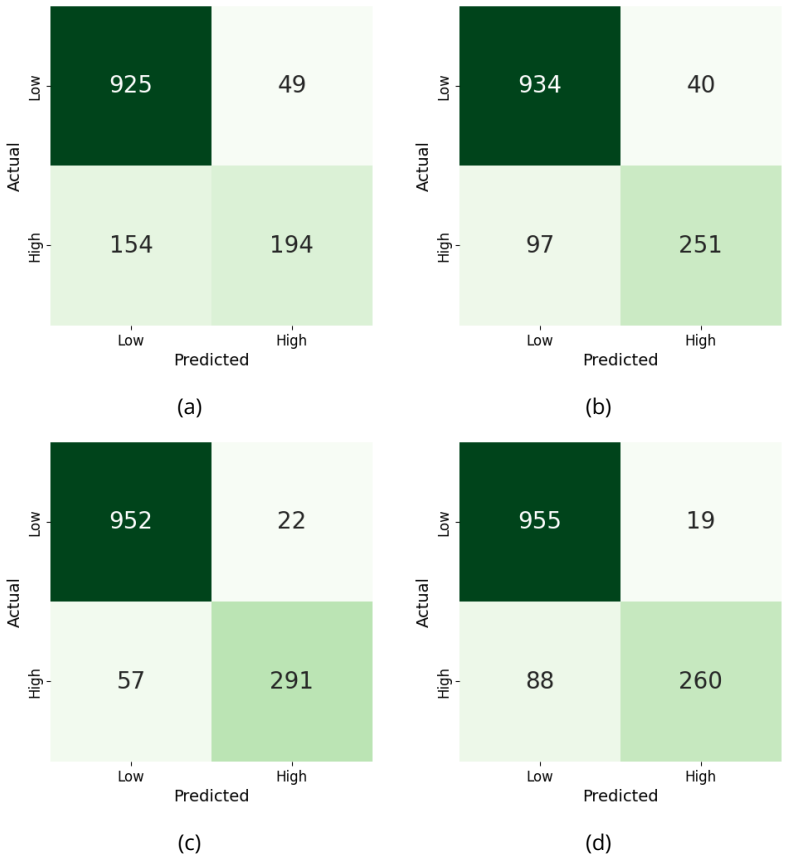

Figure 1. Confusion matrices of the four machine learning models: (a) KNN, (b) Naïve Bayes, (c) SVM, and (d) Random Forest, showing the classification results for student achievement categories.

mechanism still provided stable and accurate predictions across both categories.

Figure 2 illustrates the ROC curves for the four trained classifiers. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, while the Area Under the Curve (AUC) quantifies the overall performance of each model. As shown in the figure, KNN achieved an AUC of 0.8973, reflecting moderate predictive capability with some overlap between classes. Naïve Bayes improved upon this with an AUC of 0.9432, indicating stronger separation between low and high achievers despite its simplifying assumptions.

The SVM and Random Forest models achieved the highest AUC values, 0.9792 and 0.9723, respectively, demonstrating exceptional classification performance. The steep rise and early plateau of their curves near the top-left corner signify a high TPR with minimal false alarms. This confirms that both models were highly effective in distinguishing between achievement levels, with SVM performing slightly better overall. These results reinforce the earlier findings from accuracy and F1-scores, emphasizing that SVM and Random Forest are the most reliable and robust models for predicting student achievement in this dataset.

Figure 3 presents the PR curves for the four classifiers: KNN, Naïve Bayes, SVM, and Random Forest. The PR curve provides a detailed view of each model's ability to balance precision (the proportion of correctly predicted positive cases) and recall (the proportion of actual positives correctly identified). As shown, KNN demonstrates a steady but lower precision across varying recall values, reflecting its moderate ability to maintain accuracy as recall increases. The Naïve Bayes curve performs better but still shows a gradual decline in precision as recall rises, indicating challenges in handling feature dependencies within the dataset.

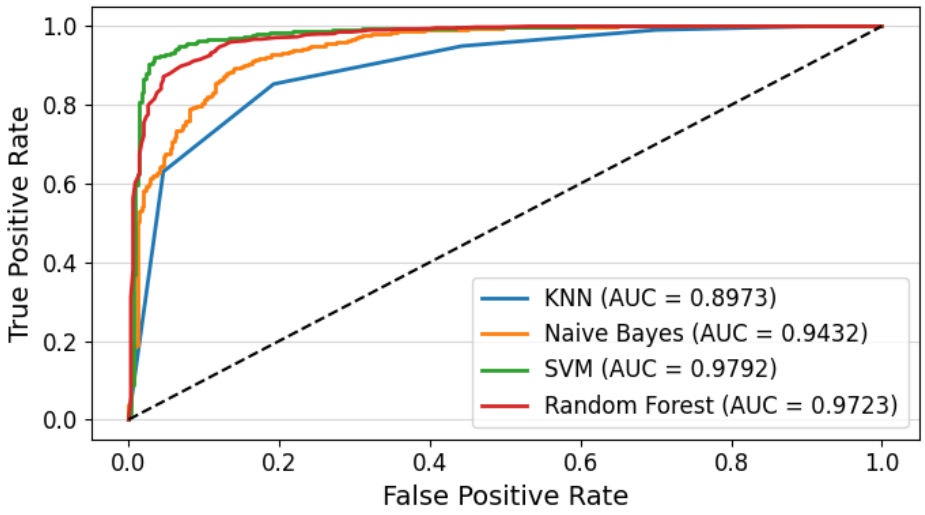

Figure 2. ROC curves showing the classification performance of KNN, Naïve Bayes, SVM, and Random Forest models.

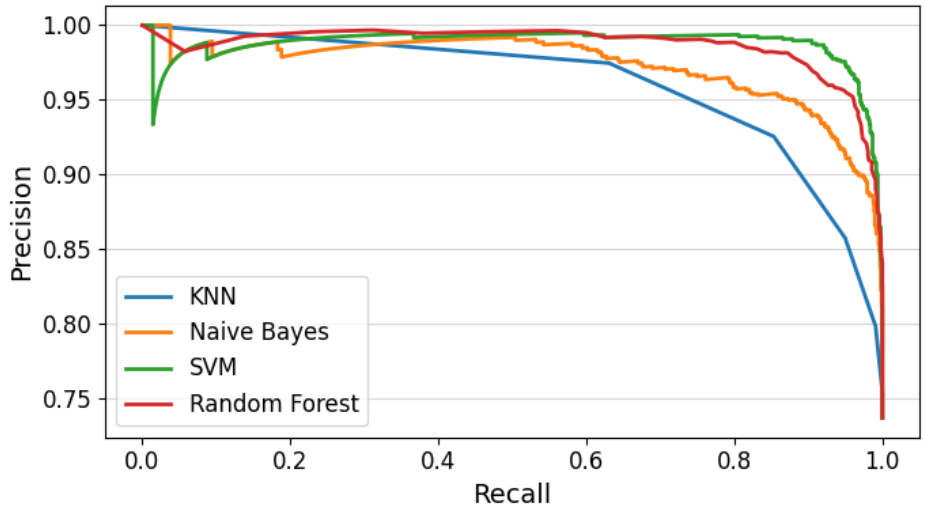

Figure 3. Precision–Recall curves comparing the predictive performance of KNN, Naïve Bayes, SVM, and Random Forest models.

In contrast, SVM and Random Forest display superior and more stable curves, maintaining high precision across nearly the entire recall range. SVM slightly outperforms Random Forest, achieving near-perfect precision at moderate recall levels, underscoring its strength in defining clear class boundaries. Random Forest closely follows, demonstrating robust precision and recall trade-offs due to its ensemble averaging, which helps mitigate overfitting. Overall, these PR curves further validate that SVM and Random Forest deliver the most reliable and consistent classification performance, effectively balancing false positives and false negatives in predicting student achievement levels.

Overall, the results demonstrate that all four machine learning models performed effectively in predicting student achievement, with varying degrees of accuracy and reliability. Among them, the SVM consistently emerged as the best-performing model, achieving the highest accuracy (94.02%), F1 Score, and AUC (0.9792), while maintaining an excellent precision–recall balance. Its ability to construct an optimal separating hyperplane allowed it to handle the multidimensional and mixed-type features of the dataset more effectively than the other classifiers. Although Random Forest also delivered strong and stable results, its performance was slightly lower than that of SVM. In contrast, Naïve Bayes and KNN exhibited solid but comparatively moderate accuracy due to their simplifying assumptions and sensitivity to noisy or overlapping data. Overall, SVM proved to be the most robust and reliable model for predicting student achievement, offering superior generalization and classification performance across all evaluation metrics.

The findings of this study carry important implications for educational practice and management. The strong

predictive performance of the machine learning models, particularly the SVM, demonstrates the potential of data-driven approaches in identifying factors that influence student achievement. By integrating variables such as study habits, parental involvement, and school environment, educational institutions can use predictive analytics to identify at-risk students and design targeted interventions proactively. For example, schools could allocate tutoring support or counseling resources based on early warning signals derived from model predictions. Furthermore, the results highlight the multifaceted nature of academic success, emphasizing that student achievement is not determined solely by cognitive ability but also by behavioral, familial, and environmental factors. For policymakers and administrators, these insights underscore the importance of holistic educational strategies that consider both in-school and out-of-school influences to enhance student outcomes.

However, this study also has several limitations. The dataset, while comprehensive, was sourced from a publicly available Kaggle repository, which may not perfectly represent the diversity of real-world educational contexts. Some features, such as motivation level, parental involvement, or teacher quality, were self-reported or qualitatively coded, introducing potential subjectivity and measurement bias. Additionally, the study used default hyperparameters for all models; further optimization might yield improved results. The binary categorization of achievement levels (low and high) simplified the problem but may have overlooked nuances among students in the middle-performance range.

Future research should address these limitations by incorporating larger and more representative datasets from diverse educational systems and regions. Employing advanced techniques such as hyperparameter tuning, deep learning architectures, or ensemble hybrid models could enhance prediction accuracy and interpretability. Moreover, integrating longitudinal data would enable the tracking of changes in student performance over time, allowing educators to understand causal relationships rather than merely correlations. Finally, future studies could explore the ethical and practical aspects of deploying predictive models in education, ensuring that such tools support equity, transparency, and informed decision-making rather than reinforcing existing disparities. By refining both the methodological and contextual aspects, future work can bridge the gap between educational data science and actionable policy development, ultimately promoting more effective and inclusive educational management practices.

4. Conclusions

This study examined the predictive capabilities of machine learning algorithms, specifically KNN, Naïve Bayes, SVM, and Random Forest, in predicting student achievement based on study habits, parental involvement, and school environment. Among the models, SVM emerged as the most effective, demonstrating the highest accuracy and balanced performance across all metrics. The findings highlight that academic success is influenced by an interplay of behavioral, familial, and institutional factors, reinforcing the need for holistic and data-informed educational strategies. By applying predictive analytics, educators and administrators can proactively identify students at risk, allocate resources more efficiently, and design targeted interventions to improve learning outcomes. Overall, this study underscores the transformative potential of machine learning in enhancing educational decision-making and promoting evidence-based management practices that support equitable and effective learning environments.

Author Contributions: Conceptualization, T.R.N. and R.I.; methodology, T.R.N. and R.I.; software, T.R.N.; validation, M.P., I.M.I. and R.I.; formal analysis, T.R.N. and M.P.; investigation, M.P. and I.M.I.; resources, M.P.; data curation, M.P. and I.M.I.; writing—original draft preparation, T.R.N. and M.P.; writing—review and editing, I.M.I. and R.I.; visualization, T.R.N.; supervision, R.I.; project administration, R.I.; funding acquisition, R.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed from Kaggle at https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Zysberg, L., and Schwabsky, N. (2021). School Climate, Academic Self-Efficacy and Student Achievement, *Educational Psychology*, Vol. 41, No. 4, 467–482. doi:10.1080/01443410.2020.1813690.
2. Hepworth, D., Littlepage, B., and Hancock, K. (2018). Factors Influencing University Student Academic Success., *Educational Research Quarterly*, Vol. 42, No. 1, 45–61.
3. Trevino, N. N., and DeFreitas, S. C. (2014). The Relationship between Intrinsic Motivation and Academic Achievement for First Generation Latino College Students, *Social Psychology of Education*, Vol. 17, No. 2, 293–306. doi:10.1007/s11218-013-9245-3.
4. Liu, J., Peng, P., Zhao, B., and Luo, L. (2022). Socioeconomic Status and Academic Achievement in Primary and Secondary

Education: A Meta-Analytic Review, *Educational Psychology Review*, Vol. 34, No. 4, 2867–2896. doi:10.1007/s10648-022-09689-y.

5. Vadivel, B., Alam, S., Nikpoo, I., and Ajanil, B. (2023). The Impact of Low Socioeconomic Background on a Child's Educational Achievements, *Education Research International*, Vol. 2023, 1–11. doi:10.1155/2023/6565088.

6. Marks, G. N., Cresswell, J., and Ainley, J. (2006). Explaining Socioeconomic Inequalities in Student Achievement: The Role of Home and School Factors, *Educational Research and Evaluation*, Vol. 12, No. 2, 105–128. doi:10.1080/13803610600587040.

7. Kyriazos, T., and Poga, M. (2024). Application of Machine Learning Models in Social Sciences: Managing Nonlinear Relationships, *Encyclopedia*, 1790–1805. doi:10.3390/encyclopedia4040118.

8. Almalawi, A., Soh, B., Li, A., and Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review, *Big Data and Cognitive Computing*. doi:10.3390/bdcc8120187.

9. Meylani, R. (2024). A Comparative Analysis of Traditional and Modern Approaches to Assessment and Evaluation in Education, *Batı Anadolu Eğitim Bilimleri Dergisi*, Vol. 15, No. 1, 520–555. doi:10.51460/baebd.1386737.

10. Cao, W., and Mai, N. (2025). Predictive Analytics for Student Success: AI-Driven Early Warning Systems and Intervention Strategies for Educational Risk Management, *Educational Research and Human Development*, Vol. 2, No. 2, 36–48.

11. Rane, N. L., Paramesha, M., Choudhary, S. P., and Rane, J. (2024). Machine Learning and Deep Learning for Big Data Analytics: A Review of Methods and Applications, *Partners Universal International Innovation Journal*, Vol. 2, No. 3, 172–197. doi:10.5281/zenodo.12271006.

12. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.

13. Noviandy, T. R., Maulana, A., Irvanizam, I., Idroes, G. M., Maulydia, N. B., Tallei, T. E., Subianto, M., and Idroes, R. (2025). Interpretable Machine Learning Approach to Predict Hepatitis C Virus NS5B Inhibitor Activity Using Voting-Based LightGBM and SHAP, *Intelligent Systems with Applications*, Vol. 25, 200481. doi:10.1016/j.iswa.2025.200481.

14. Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine Learning and Deep Learning, *Electronic Markets*, Vol. 31, No. 3, 685–695. doi:10.1007/s12525-021-00475-2.

15. Goren, O., Cohen, L., and Rubinstein, A. (2024). Early Prediction of Student Dropout in Higher Education Using Machine Learning Models, *Proceedings of the 17th International Conference on Educational Data Mining*, 349–359.

16. Namoun, A., and Alshanqiti, A. (2020). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review, *Applied Sciences*, Vol. 11, No. 1, 237. doi:10.3390/app11010237.

17. Maulana, A., Idroes, G. M., Kemala, P., Maulydia, N. B., Sasmita, N. R., Tallei, T. E., Sofyan, H., and Rusyana, A. (2023). Leveraging Artificial Intelligence to Predict Student Performance: A Comparative Machine Learning Approach, *Journal of Educational Management and Learning*, Vol. 1, No. 2, 64–70. doi:10.60084/jeml.v1i2.132.

18. Lai, N. (2025). Student Performance Factors .

19. Rahmanparast, A., Milani, M., Camci, M., Karakoyun, Y., Acikgoz, O., and Dalkilic, A. S. (2025). A Comprehensive Method for Exploratory Data Analysis and Preprocessing the ASHRAE Database for Machine Learning, *Applied Thermal Engineering*, Vol. 273, 126556. doi:10.1016/j.applthermaleng.2025.126556.

20. Ahsan, M., Mahmud, M., Saha, P., Gupta, K., and Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance, *Technologies*, Vol. 9, No. 3, 52. doi:10.3390/technologies9030052.

21. Muksalmina, M., Syahyana, A., Hidayatullah, F., Idroes, G. M., and Noviandy, T. R. (2025). Credit Card Fraud Detection Through Explainable Artificial Intelligence for Managerial Oversight, *Indatu Journal of Management and Accounting*, Vol. 3, Nos. 1 SE-Articles, 17–28. doi:10.60084/ijma.v3i1.301.

22. Fadlil, A., Herman, and Praseptian M, D. (2022). K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction, *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, Vol. 6, No. 4, 570–576. doi:10.29207/resti.v6i4.4173.

23. Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., and Ray, S. (2024). A Model-Agnostic Interpretability Approach to Predicting Customer Churn in the Telecommunications Industry, *Infolitika Journal of Data Science*, Vol. 2, No. 1, 34–44. doi:10.60084/ijds.v2i1.199.

24. Rochim, A. F., Widyaningrum, K., and Eridani, D. (2021). Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment, *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 224–228. doi:10.1109/ISRITI54043.2021.9702845.

25. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

26. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.

27. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.

28. Ferrer, L. (2022). Analysis and Comparison of Classification Metrics, *ArXiv Preprint ArXiv:2209.05355*.

29. Tharwat, A. (2021). Classification Assessment Methods, *Applied Computing and Informatics*, Vol. 17, No. 1, 168–192. doi:10.1016/j.aci.2018.08.003.

30. Cook, J., and Ramadas, V. (2020). When to Consult Precision-Recall Curves, *The Stata Journal: Promoting Communications on Statistics and Stata*, Vol. 20, No. 1, 131–148. doi:10.1177/1536867X20909693.