Excellent: International Journal of Computational Intelligence and Sustainable Innovation
Vol. 1, No. 1, June 2026
ISSN: XXXX-XXXX, DOI: 10.

CLASSIFICATION OF STUDENTS ACADEMIC ACHIEVEMENT USING A RANDOM FOREST ALGORITHM BASED ON EDUCATIONAL DATA MINING

Floricytha Sihombing[1*], Sherlyta[2], Marihot P. Parhusip[3], Kevin Frans Samuel Gultom[4]
[1234]State University of Medan, Medan, Indonesia

Article Info

Article history:

Received June 11st, 2026
Revised June 14th, 2026
Accepted June 25th, 2026

Keywords:

Educational Data Mining; Random Forest; Academic Performance Classification; Machine Learning; Student Performance Prediction;

ABSTRACT

The categorization of students' academic success presents a significant challenge owing to the effects of various academic, behavioral, and social elements that interact intricately. Precisely determining the categories of student success is crucial for facilitating educational decision-making and early intervention methods. This research sought to create and assess a model for classifying student academic performance utilizing the Random Forest technique within a framework of Educational Data Mining. A supervised machine learning approach was utilized, employing the Student Performance dataset, which comprises 2,392 records of students along with 15 attributes concerning demographic details, study patterns, parental involvement, participation in extracurricular activities, and academic results. The suggested methodology included steps such as data preprocessing, exploratory data analysis, feature selection, splitting the dataset at an 80:20 ratio, training the model, and assessing performance through accuracy, precision, recall, F1-score, analysis of the confusion matrix, evaluation of feature importance, and five-fold cross-validation. The results from the experiments indicated that the Random Forest model reached an accuracy of 90.81% with the testing dataset and exhibited robust classification results across five distinct academic achievement categories. The model performed best in GradeClass 4 and GradeClass 2, whereas lesser performance was noted in the minority classes, likely due to class imbalance. Additionally, the analysis revealed that factors related to study habits and student engagement significantly influenced the classification results. The outcomes suggested that Random Forest is an effective method for classifying multi-class academic performance and could be a dependable resource for informing data-driven educational strategies, student monitoring, and targeted academic interventions.

Corresponding Author:

Floricytha Sihombing
Email: floricytha.4233111024@mhs.unimed.ac.id

ISSN: XXXX-XXXX Doi: 10.XXX

1. INTRODUCTION

Poor academic results and the significant proportion of learners at risk of leaving school continue to be problems that have not been entirely addressed. A range of complex elements, both internal to the schools and external, impacts academic success, including hours spent studying, presence in classes, support from family, and engagement in activities outside the classroom. According to Batool et al. [1], educational institutions have gathered extensive student-related information. Nevertheless, this information is frequently not used to its full potential for making decisions based on evidence. Educational Data Mining (EDM) has arisen as a field that combines machine learning with statistical methods to uncover valuable trends from educational data to aid in delivering more effective and specific interventions. Consistent with these observations, the demand for incorporating data-driven strategies in education is increasing.

Yağcı's [2] research validates this by illustrating the dominance of Random Forest among the different algorithms that were evaluated. Random Forest continually demonstrated the highest level of precision in categorizing academic grades, signifying its status as the top algorithm for predicting educational achievements.The domain of Educational Data Mining (EDM) has swiftly developed into a unique field that merges machine learning, statistics, and data analysis to derive insights from educational datasets. A comprehensive review by Xiao et al.[3] outlines this progression, highlighting that machine learning approaches rooted in EDM particularly Decision Trees, Random Forests, and Neural Networks—are mainly employed to forecast student outcomes. This conclusion is consistent with the findings of Sarker et al.[4], which reveal that analyzing patterns in educational data can uncover significant elements that affect academic success, thus facilitating the effective utilization of EDM to enhance curricula, teaching techniques, and assessment approaches driven by data.

The Random Forest method has shown considerable success in categorizing educational data because of its capability to manage intricate characteristics, its robustness against overfitting, and its ability to produce insights that are educationally relevant regarding various features. Research by Feng et al.[5] indicated that Random Forest could reach an accuracy level of as high as 93% in predicting student performance. In addition, Kumar et al.[6] established that Random Forest is superior to XGBoost, revealing that ensemble algorithms outperform individual classification models, offering comparable accuracy alongside enhanced interpretability through the analysis of feature importance. Numerous studies indicate that academic success is affected by a range of factors. Ahmed [7] recognized academic background as a key determinant.

Moreover, Kumah et al.[8] revealed a notably strong correlation between the engagement of parents and the academic achievements of students. Rates of absenteeism and study patterns serve as significant indicators as well; Khoirunnisa & Sugiyarto.[9] along with additional studies [10] validate the impact of these factors on the categorization of student grades. In addition, indicators at the school level have also appeared as key determinants in various results derived from analyses that utilize Random Forest methods [11] [12]. Nonetheless, there remains a shortcoming in the current body of research. A majority of the studies concentrate solely on classifying into two categories (pass/fail), whereas the evaluation of performance using multi-class models that divide students into more than two grading groups is still lacking in exploration. Bujang et al.[13] have illustrated the advantages of employing the multi-class methodology.

Conversely, the challenge of class imbalance in educational datasets presents a distinct problem, which methods like SMOTE have shown to be successful in tackling. In addition, the analysis of feature importance has not been broadly adopted in real-world, actionable educational strategies.[14] This research intends to fill this void by employing the Random Forest algorithm to categorize students' academic achievements into five different grade groups (A, B, C, D, F) according to GPA, utilizing a dataset that comprises 15 multidimensional characteristics.

According to the aforementioned explanation, the research inquiries explored in this investigation are: (1) the application of the Random Forest algorithm to categorize students' GradeClass based on their learning habits, support from parents, and involvement in extracurricular activities; (2) the accuracy of the resulting model evaluated through metrics such as accuracy, precision, recall, and F1-score; and (3) the features that play the most significant role in determining students' academic success.[15] The aim of this research is to develop and assess a Random Forest classification model and to pinpoint the key factors affecting students' academic outcomes to inform data-driven recommendations for interventions.[16] This is backed by comparative research that validates the effectiveness of ensemble methods in recognizing intricate patterns within diverse academic data for multi-class applications. The uniqueness of this research is found in utilizing the Random Forest algorithm to classify students' academic achievements into five GradeClass categories through a multi-class methodology within the Educational Data Mining framework. In contrast to earlier studies, which mainly concentrated on binary classifications (such as pass or fail), this research assesses Random Forest's effectiveness in managing multi-class classification while identifying the key factors impacting academic performance via an analysis of feature importance.[17]

ISSN: XXXX-XXXX

2. METHOD

2.1 Research Design

This research uses supervised machine learning techniques to categorize students' academic outcomes into five distinct grade categories (GradeClass). [14]The choice of supervised learning was made due to the access to labeled data within the dataset, which aligns with the methodology that illustrates how this technique can identify patterns from historical data with labels. The Random Forest Classifier was selected as the main algorithm because of its capability to handle data with many features and to indicate the significance of each feature. All phases of the research were executed on the Google Colaboratory platform utilizing the Python programming language, where Python, along with the scikit-learn library, is extensively recognized as a typical tool for implementation in educational data mining research. The process consists of a number of consecutive steps: gathering the dataset, conducting exploratory data analysis (EDA), performing preprocessing, selecting targets and features, dividing the data, training the model, and assessing performance.[18]

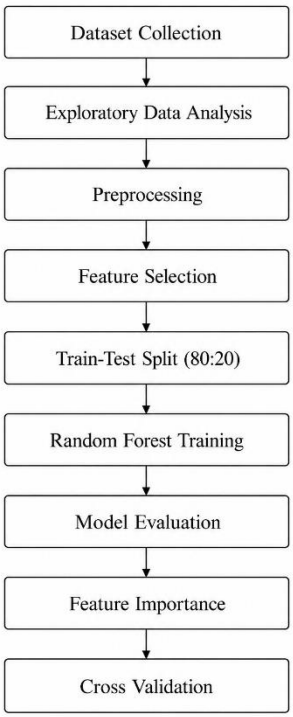

Figure 1. Research flowchart

2.2 Dataset

This research employs the Student Performance Dataset obtained from the Kaggle repository. It contains 2,392 samples of student observations, with each sample featuring 15 distinct characteristics, which include demographic information, learning habits, parental engagement, involvement in extracurricular activities, and academic performance records. The categorization of student performance relies on the target variable GradeClass, which is classified into five grade categories based on GPA. As all the individual attributes are already expressed in numerical format from the beginning, there is no need for a data encoding phase in this procedure.[19]

Table 1. Description of Features in the Student Performance Dataset

Features	Description	Type
Student ID	Unique identifier for each student	Integer
Age	Student age (15–18 years)	Integer
Gender	Gender (0=Male, 1=Female)	Nominal
Ethnicity	Student ethnicity (0=Caucasian, 1=African American, 2=Asian, 3=Other)	Nominal

ISSN: XXXX-XXXX Doi: 10.XXX

Features	Description	Type
Parental Education	Parental education level (0=None to 4=Higher)	Ordinal
Study Time Weekly	Hours of study per week (0–20 hours)	Numerik
Absences	Number of absences in a year (0–30)	Integer
Tutoring	Private tutoring status (0=No, 1=Yes	Biner
Parental Support	Parental support (0=None to 4=Very High)	Ordinal
Extracurricular	Extracurricular participation (0=No, 1=Yes)	Biner
Sports	Sports participation (0=No, 1=Yes)	Biner
Music	Music participation (0=No, 1=Yes)	Biner
Volunteering	Volunteer activity participation (0=No, 1=Yes)	Biner
Grade Class (Target)	Grade classification: 0=A (GPA ≥ 3.5), 1=B, 2=C, 3=D, 4=F (GPA < 2.0)	Ordinal

The GPA variable is excluded from being a predictive feature due to the classification of GradeClass being established according to GPA intervals. Including GPA as a feature might result in data leakage, potentially leading to an exaggerated enhancement of the model's effectiveness.[20]

2.3 Preprossesing Data

The preprocessing phase plays an essential role in influencing the model's quality. Inaccurate handling of missing data can hinder classification precision, thus making its proper management a crucial process. The data preprocessing phase involves evaluating the appropriateness of the dataset's format, identifying absent values, and verifying the existence of duplicate records. An extensive preprocessing strategy has proven to significantly enhance the effectiveness and consistency of predictive models. Additionally, statistical visualization methods are utilized to investigate class distributions and the relationships among variables. As all categorical features have been converted into numerical format, further transformation by researchers is no longer necessary.[21]

2.4 Feature Selection and Data Partitioning

In this framework, GradeClass acts as the outcome or target variable (y), whereas the other characteristics are regarded as independent or explanatory variables (X). The information is divided into 80% for training and 20% for testing, utilizing stratified sampling to preserve an even distribution among the categories. This approach of stratified sampling theoretically minimizes estimation bias by guaranteeing that each segment of data reflects identical category ratios. To make certain that this test can be conducted again with dependable (reproducible) outcomes, the random state setting is adjusted to 42.[22]

2.5 The Random Forest Algorithm

The conceptual basis of the Random Forest algorithm originates from significant sources within the field of ensemble learning studies. This method in ensemble learning functions by building numerous decision trees at the same time and then arrives at the ultimate prediction by means of majority voting. To reduce the chance of overfitting while enhancing precision, this algorithm integrates the principles of bagging along with the selection of random features. [23]

Table 2. Random Forest Parameters

Parameter	Value	Description
n_estimators	100	Number of decision trees in the forest
random_state	42	Seed for reproducibility
criterion	gini (default)	Split quality metric
max_features	sqrt (default)	Number of features considered at each split

ISSN: XXXX-XXXX

Parameter	Value	Description
bootstrap	True (default)	Use of bootstrap sampling

2.6 Model Evaluation

The performance of the algorithm was assessed through four primary metrics: Accuracy, Precision, Recall, and F1-Score. To examine the classification error rates or misclassifications among different classes, a Confusion Matrix was utilized in this research. Additionally, the analysis of feature importance was conducted to determine the attributes that most significantly influenced the prediction outcomes. At the same time, the model's capacity to generalize across data is regularly confirmed through the use of the 5-fold cross-validation method.[24]

Table 3. Evaluation Metrics

Metrik	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Proportion of correct predictions
Precision	TP / (TP + FP)	Positive prediction accuracy
Recall	TP / (TP + FN)	Ability to detect positive classes
F1-Score	2 x (Precision x Recall) / (Precision + Recall)	Average harmonic of precision and recall
Cross-Validation	k=5 fold	Cross-validation for generalization estimation

3. RESULTS AND DISCUSSION

3.1 Implementation of the (Random Forest) Model

The database explored in this study is Student Performance, which contains 2,392 student observations with a total of 15 attributes. In line with the application of artificial intelligence (AI) to predict academic achievement, GradeClass is specifically used as the target variable. The remaining 13 variables are used as predictors, following a filtering process that eliminated the StudentID attribute because it serves purely as a unique identifier.

Table 4. Data Preprocessing Table

Condition	Count Before	Count After	Notes
Missing Value	2392	2392	No missing data
Duplicate Data	0	0	No duplicate data

The preprocessing stage (data cleaning) indicates that this dataset is of very high quality. No missing values or duplicate data were found among the 2,392 data rows.

Table 5. Training and Testing Data Tables

Dataset	Number of Data Points
Training	1913
Testing	479

The entire dataset was retained to preserve data integrity before being split into training data (80% or 1,913 data points) and test data (20% or 479 data points).

ISSN: XXXX-XXXX Doi: 10.XXX

The Random Forest algorithm was selected as the AI architecture to classify student grade categories (GradeClass). The model was trained using 1,913 training data points to identify patterns in the relationship between demographic factors, family support, and academic habits and student final grades.

3.2 Evaluation Results and Discussion

3.2.1 Overall Classification Performance

Following the analysis of 479 testing data samples, this AI model attained an accuracy percentage of as much as 90.81%. This success showcases the algorithm's significant reliability in autonomously forecasting students' educational classifications. However, it remains essential to conduct a broader assessment utilizing additional evaluation criteria. This necessity arises from the fact that depending exclusively on one accuracy measure is viewed as inadequate, particularly in situations with an unequal class distribution.

Table 6. Model Accuracy Results Table

Metrik	Value
Accuracy	90,81%

3.2.2 Analysis of Precision, Recall, and the Confusion Matrix

An in-depth evaluation using the Classification Report and Confusion Matrix reveals disparities in AI performance across different student categories:

Majority Class (Grade 4 & 2): The model demonstrated optimal performance in Grade 4 (F1-Score 0.96), successfully predicting 232 data points accurately. This is consistent with Grade 2 (F1-Score 0.93). This high level of precision is due to the abundance of data representation in these grade groups within the dataset, providing the AI with ample examples to learn from.

Table 7. Precision, Recall, and Confusion Matrix Table

GradeClass	Precision	Recall	F1-Score
0	0.67	0.29	0.40
1	0.78	0.91	0.84
2	0.90	0.95	0.93
3	0.88	0.89	0.89
4	0.96	0.95	0.96

Minority Class (GradeClass 0): The students in Class 0 are those whose traits are the hardest for the model to recognize, demonstrated by the F1-Score dropping to 0.40. Among 21 real data points from Class 0, the model managed to accurately predict just 6. This challenge often arises in AI applications, as Class 0 probably corresponds to unusual cases or a smaller group of students (such as high achievers) for whom there isn't enough information for the model to detect trends.

3.2.3 Confusion Matrix Analysis

The distribution of the model's correct and incorrect predictions is visualized in the Confusion Matrix:

Table 8. Confusion Matrix Table

Current\Predictions	0	1	2	3	4
0	6	6	3	2	4

ISSN: XXXX-XXXX

1	1	49	2	1	1
2	0	0	74	3	1
3	0	4	2	74	3
4	2	4	1	4	232

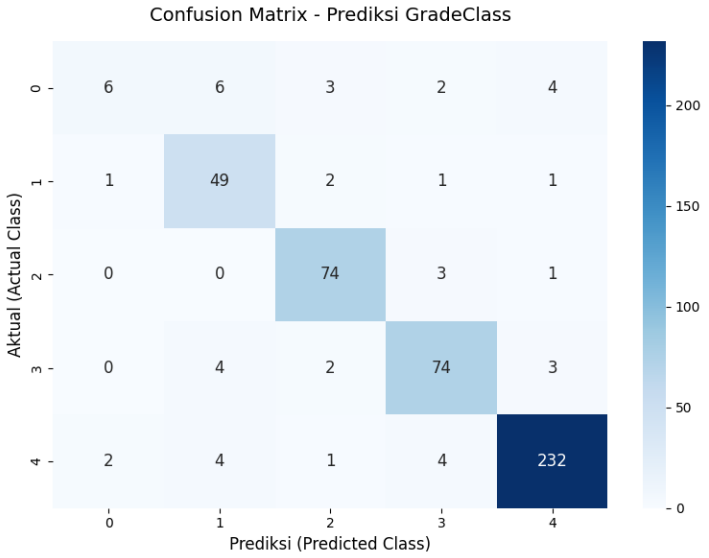

Figure 2. Confusion Matrix

3.2.4 Analysis of Performance Determinants (Feature Importance)

An analysis of feature significance offers crucial educational insights regarding the key factors affecting student success based on the assessment of the algorithm:

GPA (Weight 49.7%): The prominence of this element is expected and logical, considering that the establishment of GradeClass benchmarks is primarily founded on groupings of GPA scores.

Absence Rate (Weight 21.1%): This observation is particularly significant within the realm of educational practices. The AI indicates that the rate of absences stands out as the non-exam measure that most significantly influences which GradeClass students belong to. A higher rate of absences strengthens the suggestion of worsening student outcomes in the evaluation carried out by the model.

Weekly Study Time & Parental Support (Weights 5.5% & 2.6%): The attributes StudyTimeWeekly and ParentalSupport function as additional influencing factors. While their significance is much less than that of the absence rate, the algorithm still considers the regularity of study time and the influence of parental assistance as credible secondary indicators in classifying students.

Table 9. Feature Importance Table

Variable	Importance
Absences	0.211184
Student ID	0.096333
Study Time Weekly	0.055345
Parental Support	0.026120

... (another variable)	...

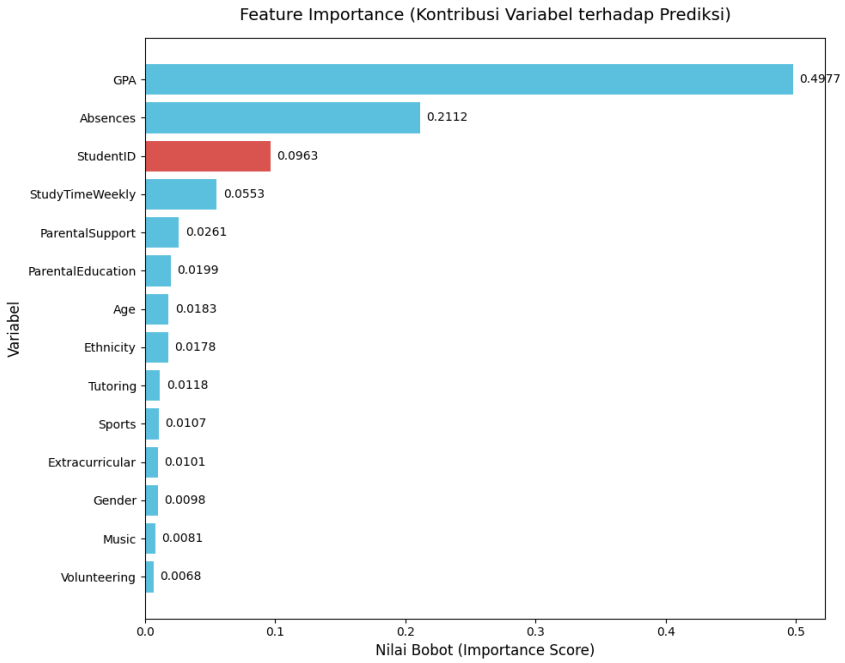

Figure 3. Feature Importance

3.3 Model Stability (Cross-Validation)

Cross-validation evaluations employing 5-fold cross-validation produced a mean accuracy of 87.62%. While this average is quite impressive, a significant variation was observed in Fold 5, which experienced a decline in accuracy to 50.63%. This situation indicates that there is a class imbalance within the dataset. When a data fold happens to randomly exclude instances from particular classes (like Class 0 or Class 1), the AI's capacity for generalization diminishes, resulting in a sharp decrease in its accuracy.

Table 10. Cross Validation Table

Fold	Accuracy
1	96,87%
2	98,12%
3	97,70%
4	94,77%
5	50,63%

4. CONCLUSION

According to the findings of the study, the Random Forest algorithm has effectively classified students' academic achievements into five distinct grade levels (A, B, C, D, and F) by considering various factors including study routines, parental involvement, and student engagement in activities. This model was created utilizing the Student Performance dataset, which contains 2,392 entries of student information.

An assessment of the outcomes indicates that this model exhibits strong performance, achieving an accuracy rate of 90.81% when evaluated against the test set. Furthermore, the metrics for precision, recall, and F1-score demonstrate that the model can accurately classify the majority of categories, especially GradeClass 2

ISSN: XXXX-XXXX

and GradeClass 4. Hence, the goal of developing and assessing the Random Forest classification model has been successfully met.

Analyzing the importance of features indicated that GPA stands out as the most influential element in the classification, followed by factors such as Absences, Weekly Study Time, and Parental Support. These findings reinforce the idea that attendance, study time, and parental involvement are crucial aspects that the model uses to assess and classify students' academic performance. As a result, the aim of identifying the key factors impacting students' academic success has been fulfilled.

Future studies could enhance the model's capabilities by incorporating techniques to address data imbalance, like SMOTE, engaging in hyperparameter tuning, and evaluating the effectiveness of Random Forest in comparison to other algorithms such as XGBoost, LightGBM, and CatBoost to better classify students' academic performance.[25]

ACKNOWLEDGMENTS

The author would like to express gratitude to all team members who collaborated and contributed to the preparation and completion of this scientific paper. The author also deeply appreciates the instructor of this course, who provided guidance, direction, feedback, and support throughout the research process and the writing of this article. It is hoped that this research will be useful for the advancement of the field of science and serve as a reference for future research.

REFERENCES

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol. (Dordr).*, vol. 28, no. 1, pp. 905–971, 2023.

[2] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[3] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Engineering Reports*, vol. 4, no. 5, p. e12482, 2022.

[4] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, "Analyzing students' academic performance using educational data mining," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100263, 2024.

[5] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022.

[6] M. Kumar, N. Singh, J. Wadhwa, P. Singh, G. Kumar, and A. Qtaishat, "Utilizing random forest and XGBoost data mining algorithms for anticipating students' academic performance," *International Journal of Modern Education and Computer Science*, vol. 16, no. 2, pp. 29–44, 2024.

[7] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied computational intelligence and soft computing*, vol. 2024, no. 1, p. 4067721, 2024.

[8] P. K. Kumah, S. T. Baidoo, and H. Yusif, "Investigating the role of parental involvement in enhancing academic performance of tertiary students: evidence from the Kwame Nkrumah University of Science and Technology, Kumasi," *Cogent Education*, vol. 11, no. 1, p. 2361997, 2024.

[9] A. A. P. Sari and A. Buchori, "Penerapan Model Problem Based Learning Untuk Meningkatkan Kemampuan Pemecahan Masalah Matematis Siswa SMA Pada Materi SPLTV," *Supermat: Jurnal Pendidikan Matematika*, vol. 8, no. 1, pp. 28–43, 2024.

[10] Y. P. C. Dewi, A. Anas, and L. Lutfiyah, "The Influence of PBL Learning Model on High School Students' Learning Outcomes in System of Linear Equations in Three Variables Material," *ETDC: Indonesian Journal of Research and Educational Review*, vol. 4, no. 3, pp. 778–788, 2025.

[11] S. Tosun and D. B. Kalaycıoğlu, "Data mining approach for prediction of academic success in open and distance education," *Journal of Educational Technology and Online Learning*, vol. 7, no. 2, pp. 168–176, 2024.

[12] R. Tertulino and R. Almeida, "A Multi-level Analysis of Factors Associated with Student Performance: A Machine Learning Approach to the SAEB Microdata," *arXiv preprint arXiv:2510.22266*, 2025.

[13] S. D. A. Bujang *et al.*, "Multiclass prediction model for student grade prediction using machine learning," *Ieee Access*, vol. 9, pp. 95608–95621, 2021.

[14] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 2, 2024.

[15] J. Wang and Y. Yu, "Machine learning approach to student performance prediction of online learning," *PLoS One*, vol. 20, no. 1, p. e0299018, 2025.

[16] A. Palanivinayagam and R. Damaševičius, "Effective handling of missing values in datasets for classification using machine learning methods," *Information*, vol. 14, no. 2, p. 92, 2023.
[17] R. M. Kalita and S. Baruah, "Data Preprocessing and Missing Data Handling for Predicting High School Academic Outcomes," *INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES*, pp. 1762–1768, 2026.
[18] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, p. e2245, 2024.
[19] O. S. Kalange, R. S. Kahat, A. S. Kale, T. R. Kale, and P. S. Joglekar, "Implementation of Various Machine Learning Algorithms for Traffic Sign Detection and Recognition," 2022.
[20] E. A. Yassine, K. Mohammed, and J. Youness, "Mathematical Modeling of Monetary Poverty by K-Nearest Neighbors Algorithm," in *The International Conference on Artificial Intelligence and Smart Environment*, Springer, 2024, pp. 190–195.
[21] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, p. 268, 2025.
[22] V. W. Lumumba, D. Kiprotich, M. Lemasulani Mpaine, N. Grace Makena, and M. Daniel Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," *K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models (June 01, 2024)*, 2024.
[23] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset," *Electronics (Basel).*, vol. 12, no. 21, p. 4423, 2023.
[24] Z.-H. Geng, Y. Zhu, P.-Y. Fu, Y.-F. Qu, Q.-L. Li, and P.-H. Zhou, "A comparative analysis of prognostic regression models and machine learning algorithms in surgical decision-making of cardial submucosal tumors," *Gastroenterology & Endoscopy*, vol. 2, no. 1, pp. 19–24, 2024.
[25] M. Thahiruddin, S. Khotijah, A. El Farras, and A. I. Hasan, "A Comparative Analysis of Deep Learning Architectures for The Classification of Madura Sliced Tobacco," *Jurnal Teknologi dan Open Source*, vol. 9, no. 1, pp. 37–47, 2026.