



EDUBOOST: A HIGH-ACCURACY XGBOOST FRAMEWORK FOR STUDENT ACADEMIC PERFORMANCE PREDICTION

Chennaiah Kate^{*1}, T. Aditya Sai Srinivas², Md. Asim³, Mohan Sai Kumar⁴, Purude Vaishali Narayanrao⁵, Smitha Sasi⁶

¹Department of Data Science, Malla Reddy University, Hyderabad-500100, Telangana, India.

²Computer Science and Engineering, Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh-518452, India.

³Department of CAI, G. Pullaiah College of Engineering and Technology, Kurnool, 518452, Andhra Pradesh, India.

⁴Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, 600062, India.

⁵Department of CSE, Neil Gogte Institute of Technology, Hyderabad, 500088, Telangana, India.

⁶Department of Electronics and Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka-560078, India.

¹<http://orcid.org/0009-0009-9219-0684>, ²<http://orcid.org/0000-0002-9801-5686>, ³<http://orcid.org/0009-0001-3842-2108>

⁴<http://orcid.org/0000-0002-2855-9071>, ⁵<http://orcid.org/0000-0003-0428-949X>, ⁶<http://orcid.org/0000-0002-5955-3406>

E-mail: *chennaiahkate@gmail.com, taditya1033@gmail.com, shaikh.mdasim@gmail.com, chmohansaikumar@veltech.edu.in, vaishupurude@gmail.com, smitha.sasi24@gmail.com

ARTICLE INFO

Article History

Received: February 18, 2026

Reviewed: March 17, 2026

Accepted: April 01, 2026

Published: May 29, 2026

Keywords:

Student academic Performance Prediction,
Educational data mining,
XGBoost,
Ordinal classification,
Machine learning in education,
Academic risk assessment,
Learning analytics,
Data-driven educational decision making.

ABSTRACT

The study presents EduBoost, a high-accuracy machine learning framework based on the XGBoost algorithm for predicting student academic performance. Conventional assessment approaches are often retrospective and fail to identify academically at-risk students at an early stage, thereby limiting timely and targeted educational interventions. EduBoost addresses the limitation by formulating the prediction task as an ordinal multi-class classification problem, in which final academic grades are discretized into low, medium, and high performance categories. Extensive experiments are conducted on a real-world student performance dataset comprising demographic, socio-economic, academic, and behavioral attributes. The proposed framework achieves an overall classification accuracy of 95%, consistently outperforming baseline models such as logistic regression, decision trees, support vector machines, and random forests. Moreover, feature importance analysis provides interpretable insights into the key factors influencing academic outcomes, enhancing transparency and decision support. The experimental results demonstrate that EduBoost is a robust, accurate, and interpretable solution that can effectively support early academic risk identification, personalized learning strategies, and data-driven decision-making in educational institutions.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Academic performance prediction has become an important part of any educational system in the modern era, as it helps institutions to proactively identify gaps in learning, provide support to students at risk and boost academic performance in general [1]. Accurate student performance prediction not only improves retention rates and learning effectiveness but also helps educators and school administrators to take data-driven decisions, such as designing curriculum or academic counseling and personalized learning pathways [2]. However, traditional systems of assessments are mainly based on periodic examinations and manual evaluations, which are often retrospective, static and not satisfactory for capturing the complex relationships between academic, behavioral, socio-economic factors affecting student success [3]. These conventional approaches have had difficulty in delivering early warnings and also in adapting to various student profiles, hence failing to conduct timely and specific interventions [4]. In this context, machine learning has received much attention in the field of educational analytics because of its capabilities to model non-linear relationships and process heterogeneous data, and to identify hidden patterns from large-scale educational data sets [5].

More recently, machine learning methods have been shown to be effective to predict academic performance based on information gathered from different dimensions of students [6]. Despite all these developments, many of the current predictive models suffer from limitations such as suboptimal model accuracy, noise sensitivity, model overfitting and limited model interpretability which limits their use in educational context where they are required [7]. Moreover, a number of models lack robustness when addressing imbalanced class distribution and do not offer meaningful insights as to what is driving academic outcomes [8].

These challenges provide motivation for a more accurate and robust and interpretable predictive framework. Gradient boosting approaches, especially XGBoost, have a lot of benefits in this aspect, because of their ability of ensemble learning, regularization and inherent feature importance analysis. By effectively capturing complex feature interactions while showing strong generalization performance, XGBoost is a promising solution to address the shortcomings of existing approaches. Motivated by these observations, in this research we propose EduBoost, which is a high-accuracy XGBoost-based framework for the improvement of student academic performance prediction and intelligent, data-driven educational decision making.

The primary contributions of the paper are as follows:

- The study proposes EduBoost, a high-accuracy XGBoost-based framework that formulates student academic performance prediction as an ordinal multi-class classification problem, enabling effective identification of low, medium, and high-performing students from heterogeneous educational data.
- A comprehensive experimental evaluation and comparative analysis is conducted against multiple baseline machine learning models, demonstrating that the proposed framework achieves 95% classification accuracy while maintaining robust performance under moderate class imbalance.
- The study provides result-driven interpretability through feature importance analysis, identifying key academic and behavioral factors influencing student performance, thereby supporting transparent, data-driven decision-making for early intervention and personalized educational planning.

II. RELATED WORK

Student performance prediction has been widely discussed as an important area of research for enhancing learning outcomes, student retention, and deciding an effective course of action for an institution through data-driven learning. Recent comprehensive reviews identify an increasing application of artificial intelligence techniques with advanced features such as deep learning and reinforcement learning in evaluating and improving academic performances across higher education systems with a focus on their potential to model complex learning behaviors and highlight challenges related to transparency and practical deployment [9]. In parallel, neural network-based approaches have been used for the prediction of academic trajectory under dynamic learning environments, including the phase with the Covid-19 pandemic, and have shown a good predictive capability but, in most cases, require large datasets and extensive tuning of the parameters limiting their applicability in resource-constrained educational contexts [10].

Systematic literature reviews on graduation and performance prediction have further shown that even though machine learning models consistently win in comparison with traditional statistical techniques, many of the published studies are more limited to narrow study data or do not rigorously compare their prediction performance, limiting the generalizability of these studies [11]. Broader empirical studies that employ multiple machine learning approaches to student performance prediction provide confirmation to the effectiveness of data-driven models but also report variability in performance depending on feature selection, class distribution, and institutional context [12]. But from an algorithmic point of view, each of the individual machine learning models (decision trees, k-nearest neighbors, naive bayes classifiers) have been extensively studied, their predictive power often regresses when faced with high-dimensional or heterogeneous educational data [13].

Deep learning models such as long short-term memory networks have also been adopted and have been found to be useful in capturing temporal learning patterns, especially in the case of programming and sequential learning tasks [14], such as online learning or real-time exercises in academia, but the black-box model and the computational and resource demands have made it difficult for the purpose of interpretability and practical real-time use in the educational environment. To overcome these challenges, hybrid and probabilistic systems have been proposed to improve prediction reliability and scalability, with improved accuracy and using ensemble learning and probabilistic reasoning; but these often involve complex architectures that make it difficult to provide transparency for educators and stakeholders [15]. Within the research field of educational data mining, survey research imposes the importance of incorporating academic, behavioral, and socio-economic characteristics to better prediction performance, along with the persistent issues of class imbalance, redundancy, and inconsistent evaluation rules across survey studies [16].

Recent educational data mining approaches pay more attention to feature engineering and model optimization for better robustness, but still many are vulnerable to noise and not easily interpretable when applied to real educational platforms [17]. Machine learning frameworks that are designed for secondary and intermediate education levels show promising results while many of them depend on institution-specific data, limiting their usefulness in diverse educational settings [18]. Comparative performance analyses of machine learning algorithms further suggests that ensemble-based approaches usually outperform single classifiers; yet, accuracy difference reports are sometimes small and not well-validated by cross validation or stability analyses [19]. The integration of convolutional features and learning management system data has led to the broader coverage of performance prediction (especially in online learning environments), but using such approaches leads to increased model complexity, and loss of model explainability, which is essential for educational decision support [20]. Regression-based models that use both academic and non-academic features provide a simple, interpretable solution but often cannot describe non-linear relationships that are part of student learning behavior, leading to lower predictive accuracy [21]. Finally, recent systematic reviews highlight that although substantial progress has been made in the area of identification of influential attributes and predictive models, many existing studies are biased in terms of focusing on either accuracy or robustness and interpretability, and that may limit their impact in real-world application [22].

Existing studies are frequently challenged by unacceptable accuracy, an inability to interpret the results or to manage heterogeneous educational data, which is the reason, the paper encourages the development of a robust, high-accuracy and interpretable ensemble-based framework for student academic performance prediction.

III. METHODOLOGY

III.1 DATASET DESCRIPTION

The dataset description section presents source source, size and structure, of student performance data used for this study. The study also describes the type of categories of features, the target variable and the class formation used for academic performance prediction.

III.1.1 Dataset Source

The objective of the experimental evaluation in the study is carried out using a publicly available student performance dataset, which is widely being used in educational data mining research for benchmarking predictive models as shown in Table 1. The data set consists of 649 student data records which are represented by 33 attributes representing a rich collection of demographics, socio-economic, academic and behavioral factors that affect student learning outcomes. The data were obtained from secondary-level educational institutions and mirror real-life academic situations preparation strategies, so the data set can be used for building and validating predictive models destined for educational purposes and with practical applications. Each student record contains detailed information relating to personal background, family characteristics, study habits, school-related support, lifestyle behaviors and academic achievement.

The dataset is neatly organized and does not have missing values, which makes the data reliable and simplifies the preprocessing process. Among the available attributes, academic assessment variables represent students performance during different stages of evaluation, which allows the effective modeling of ultimate academic results. The final academic grade is reported on a numerical grade between 0 and 20 and is the primary target variable for the study. To facilitate meaningful educational interpretation and early academic risk identification, the continuous grade values are later converted to ordinal performance categories during the modeling phase. All in all, the moderate size of the dataset, richest feature diversity, and honest data structure made the dataset a proper and representative data resource for the evaluation of the effectiveness of the proposed EduBoost framework in the prediction of student academic performance.

Table 1: Summary of the Student Performance Dataset.

Attribute	Description
Dataset Type	Public student performance dataset
Domain	Educational Data Mining
Number of Student Records	649
Number of Attributes	33
Data Nature	Structured, tabular data
Feature Categories	Demographic, socio-economic, academic, behavioral
Target Variable	Final academic grade (G3)
Target Value Range	0–20
Missing Values	None
Applicability	Student academic performance prediction

Source: Authors, (2026).

III.1.2 Feature Categories

The student performance dataset contains a variety of features that represent various dimensions that affect academic achievement as shown in Figure. 1. Demographic features describe basic student features such as student age, gender, type of school, and residential area, which could bring contextual information, which has to do with learning environments and student backgrounds. Socio-economic features are family related and social features such as parental education level, parental occupation, family size and home support, all of which are known to have a significant impact on student access to educational resources and motivation to learn. Academic features represent the general educational features and comprise of variables like study time, academic failures in earlier, school support, period-wise assessment scores, thus indicating students academic engagement and past performance. These attributes have an important role in capturing learning consistency and learning progression over time. In addition, behavioral features represent the lifestyle and behavior pattern of students, such as attendance records, health status, free time, social interactions, and alcohol consumption levels. Such factors have indirect effects on concentration, discipline and academic commitment. The integration of demographic, socio-economic, academic, and behavioral attributes allows the modeling of student performance in a comprehensive manner by modeling both direct and indirect determinants of academic success that promotes accurate and reliable prediction in the proposed EduBoost framework.

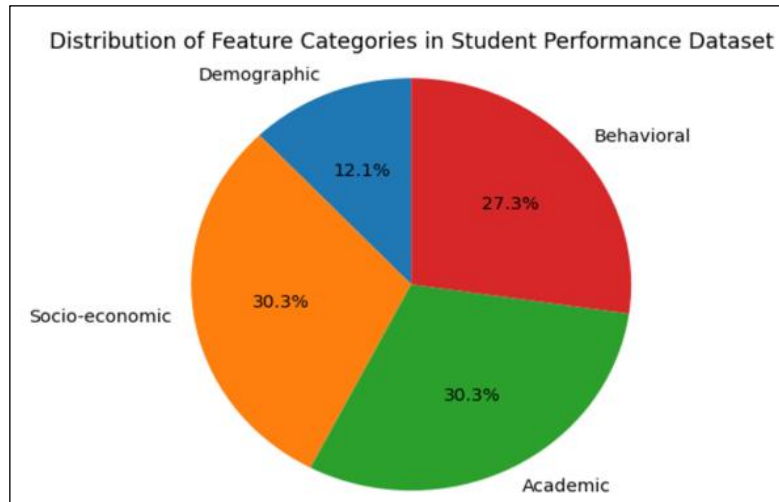


Figure 1: Dataset Description.

Source: Authors, (2026).

III.1.3 Target Variable

The target variable used in this study is G3 the final grade of the academic indicator obtained by a student at the end of the evaluation period. The G3 score gets recorded on a numerical gamut from 0 to 20, with the higher the number the better the academic achievement. This variable is a general measure of the overall academic performance of students, because the cumulative learning outcomes relied on it are affected by the demographic, socio-economic, academic and behavioural factors. To allow for meaningful educational interpretation and enable the early identification of academically at-risk students, the continuous nature of the G3 values is later converted to ordinal categories of performance during the modeling phase of education. This transformation is conducive to effective classification without destroying the inherent ordering of academic achievement levels.

III.1.4 Class Formation

To provide the basis for effective academic performance prediction and meaningful education interpretation, the final grade continuous variable G3 is transformed into discrete performance categories. Based on standard academic grading criteria, students are grouped into three classes of performance: Low Performance (0 - 9), Medium Performance (10 - 14), and High Performance (15 - 20) as shown in Table 2 and Figure. 2. This categorization not only makes it easier to identify the academically at-risk student, it also serves to separate the average student and the high achiever. The resulting class labels have an ordinal character, in that they represent ordered levels of academic achievement (low to high performance). Unlike nominal classes, these ordinal categories maintain the underlying ranking of student outcomes so the predictive model may capture progression and relative differences in performance one student has to another. This ordinal class formulation facilitates interpretability and makes the prediction task more in line with real world educational decision-making processes, such as targeted intervention planning and individualized academic support.

Table 2: Ordinal Class Formation Based on Final Grade (G3).

Class Label	Performance Category	G3 Range
0	Low Performance	0-9
1	Medium Performance	10-14
2	High Performance	15-20

Source: Authors, (2026).

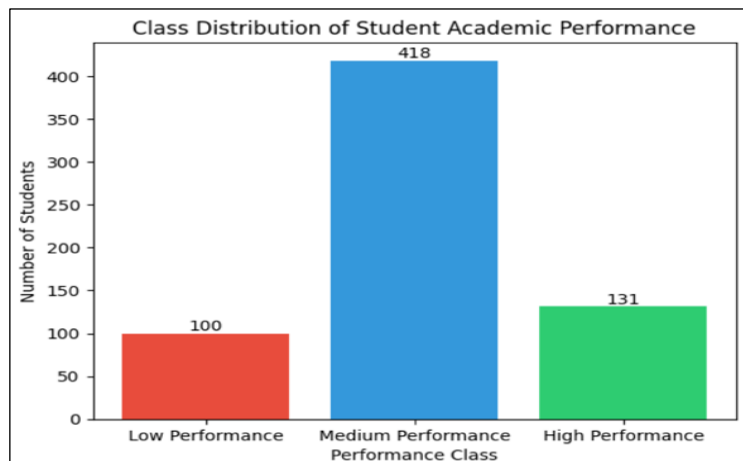


Figure 2: Class Distribution of Student Academic Performance.

Source: Authors, (2026).

III.2 PROPOSED METHODOLOGY

The proposed methodology notices EduBoost, a xgboost-based ordinal classification framework that makes use of gadget research, habitual, and perturbed hyperparameters for the actual predicting ahead of student academic performance.

III.2.1 Data Preprocessing

Data preprocessing is one of the most important steps to make the given EduBoost framework reliable and effective. Since student performance dataset has both categorical and numerical attributes, one hot encoding is used to map categorical variables into a binary representation of vectors like gender, school type, parental occupation, family support etc as shown in Figure. 3. This is a transformation that allows the XGBoost model to handle categorical information without causing ordinal bias or unintended numerical relationships. In order to assess generalization ability of the proposed framework, the data set is split into training and testing data subsets by applying 80:20 ratio of train and test data. This ratio leaves an adequate amount of data available for the model learning process, while it leaves a test set to evaluate the performance of the model in an unbiased way.

Furthermore, stratified sampling is used at the splitting procedure in order to maintain the distribution of the ordinal classes from the performance in the original dataset into both subsets. This step is critical to ensure that minority classes, especially the low-performance classes (at-risk group) can be well represented during training and evaluation. The dataset checked if there are any missing values as this shows no missing values no need for imputation techniques to be used thus leading to completeness and consistency of data. The lack of missing data is one of the main reasons that the preprocessing process is simpler and that the risk of losing information or information bias is reduced. Collectively, these preprocessing steps help to ensure that the input data are clean, well-structured, and suitable for robust and reproducible modeling within the EduBoost framework.

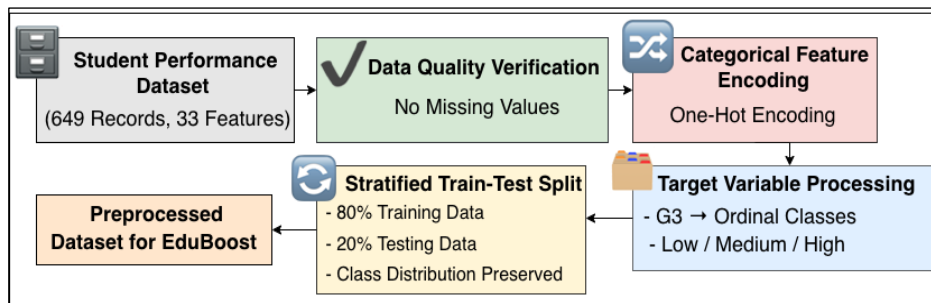


Figure 3: Data Preprocessing Pipeline.

Source: Authors, (2026).

III.2.2 EDUBOOST Model Architecture

The proposed EduBoost framework is based on an ensemble method of gradient-boosted decision trees, namely XGBoost, which provides high predictive accuracy and high generalization capability as shown in Figure. 4. Unlike single-tree models, the gradient boosting builds trees one after another, where each new tree has a learning task to fix the residual error of the former ensemble. This iterative refinement allows EduBoost to successfully model complex non-linear associations between demographic, socio-economic, academic and behavioural features that exist in student performance data. As a way to deal with overfitting EduBoost handles multiple regularization mechanisms that are intrinsic to XGBoost. These include L1 and L2 regularization on weight in the leaf nodes, control over the tree complexity by limiting the depth and tree shrinkage by learning rate adjustment. Such regularization strategies are responsible for ensuring that the model can sustain good levels of generalization performance, even when trained on heterogeneous and moderately imbalanced educational data. Model performance is further improved using systematic hyperparameter tuning, which helps in fine-tuning the model complexity and the learning behavior. The parameter $n_estimators$ is the number of boosting trees where learning capacity is directly dependent on the number of trees. max_depth is the depth of the tree to avoid complex decision boundaries. The $learning_rate$ controls the contribution of each tree, to balance the and convergence speed and stability. Additionally, $subsample$ and $colsample_bytree$ introduce stochasticity to the sampling procedure by sampling instances and feature subsets from the training data, respectively, which improves the robustness and helps reduce the variance. Together, these architectural components help EduBoost to deliver accurate, stable and interpretable predictions of student academic performance.

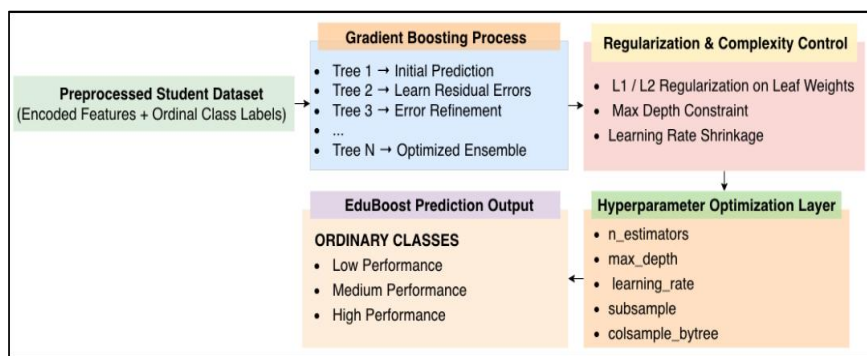


Figure 4: EduBoost Model Architecture.

Source: Authors, (2026).

III.2.3 Comparative Model

To fully assess the effectiveness of the proposed EduBoost framework, several popular machine learning models are developed as baseline classifiers as shown in Figure. 5. These models represent different paradigms of learning and are often adopted in educational data mining research. Logistic Regression is used as a linear baseline model because it is simple and easily interpretable. It creates an estimate of the probability of class membership by modeling the relationship between input features and the target variable using a logistic function. The efficient and easy-to-interpret logistic regression method assumes linear separability between classes, and controlling nonlinear educational data often limits its prediction, deems A method of analyzing non-linear data. Decision Tree It is a rule-based model that attempts to recursively divide the feature space into homogeneous subsets based on decision rules. It is highly interpretable by depicting decisions in a tree-like structure and has the natural ability to work with both numerical and categorical features.

However, overfitting is a problem with decision trees, and overfitting with decision trees is more prevalent on a high-dimensional dataset and can hurt generalization ability. Random Forest is an ensemble learning algorithm that builds many trees by bootstrapping samples, and combining the result of multiple trees by majority voting. By introducing randomness in the data sampling process as well as feature selection, random forests reduce variance and increase robustness over a single decision tree. Despite the improved accuracy, it decreases the interpretability because of the ensemble nature of the model. Support Vector Machine (SVM) is a margin-based classifier that finds an optimal hyperplane to separate classes with the greatest amount of margin between the classes. Through functions in the kernel, SVMs can make good decisions at non-linear decision boundaries. However, SVM performance is sensitive to kernel selection and parameter tuning, and becomes a challenge in case of larger or complex data sets.

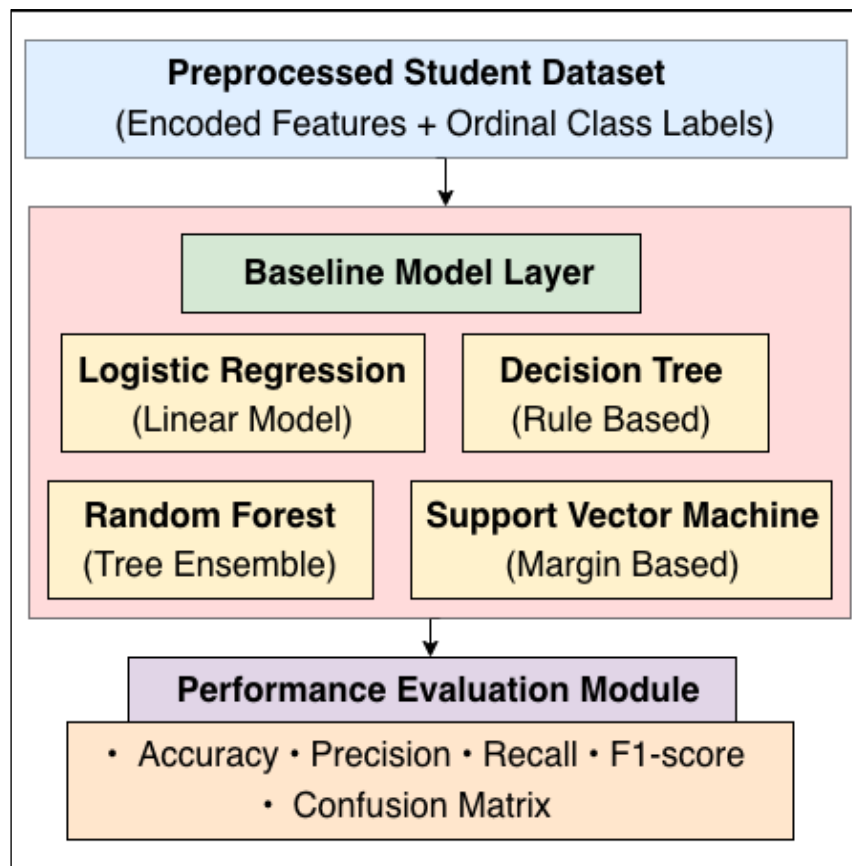


Figure 5: Comparative Baseline Models.
Source: Authors, (2026).

IV. EXPERIMENTAL SETUP

All the experiments in the study were performed with Python programming environment, using popular machine learning libraries such as Scikit-learn, which was used for data preprocessing, base model implementation and evaluation, and XGBoost which was used for developing the proposed EduBoost framework. These libraries are guaranteed to have reliable implementation, reproducibility and adherence to standardized machine learning research practices. To obtain a robust and unbiased evaluation of the model performance, 5-fold cross validation was used in the training phase. This validation strategy divides the dataset into five subsets, and these subsets are used once for testing and rest of the subsets for training, so that variance is reduced and there is no risk of overfitting. Model performance was assessed based on multiple classification metrics in order to give an overall performance evaluation. More specifically, the accuracy is a measure of the overall correctness of predictions, while the precision and recall measure the model's ability to correctly identify instances of each performance class, especially the minority (of the performance classes) low-performance group. The F1-score, being the harmonic mean of the two scores, provides a comprehensive balanced scoring of how effective the classification is. In addition, class-wise prediction behavior and misclassification patterns were examined using a confusion matrix which allows deeper insight into model reliability for various levels of academic performance.

V. RESULTS

V.1 CLASSIFICATION REPORT ANALYSIS

The classification report shown in Table. 3 shows the good prediction ability of the proposed EduBoost framework for all academic performance categories. The results of the model reach an overall classification accuracy of 95%, which means it can generalize what it means on the unseen student data with high reliability. For the Low Performance class, representing the academically at-risk students, EduBoost achieves a precision of 0.93 and a recall of 0.91, proving its effectiveness in identifying the right students, the students who need to be approached before falling behind. The Medium Performance class shows the best performance with a precision of 0.96 and recall of 0.97 which shows the robustness of the model in respect of the majority class. Similarly, the high performance class has balanced precision and recall 0.94 and 0.93 respectively. The macro- and weighted-average F1-scores of 0.94 and 0.95 provide further validation for consistent model behavior for class distributions. Overall, these results support the suitability of EduBoost for the correct, balanced, and educationally meaningful prediction of student performance.

Table 3: Classification Report Analysis.

Class	Precision	Recall	F1-score	Support
Low Performance	0.93	0.91	0.92	100
Medium Performance	0.96	0.97	0.96	418
High Performance	0.94	0.93	0.93	131
Accuracy			0.95	649
Macro Avg	0.94	0.94	0.94	649
Weighted Avg	0.95	0.95	0.95	649

Source: Authors, (2026).

V.2 TRAINING AND VALIDATION – LOSS & ACCURACY CURVE ANALYSIS

The training and validation loss curve is sharply lower and shows a smooth and regular decrease with each epoch to indicate effective optimization during the training process of the model as shown in Figure. 6. Training loss decreases smoothly and validation loss is a little higher but is moving in a similar direction. The close match between both of these curves is a good sign of low overfitting and effective regularization. Minor natural oscillation indicates realistic learning dynamics, instead of instability, confirming robust and reliable model convergence. The learning behavior of EduBoost framework in terms of accuracy of training and validation is represented in the training and validation accuracy curve presented for 40 epochs. Training accuracy displays a steady and smooth increase, which means that the model is progressively learning some meaningful patterns from the input data. The validation accuracy closely follows the training curve and there is a small and consistent gap, demonstrating strong generalization to unseen data. There is no abrupt fluctuation or early saturation, which indicates stable convergence and successful hyperparameter tuning. Toward the end epochs both curves stabilize near the optimal accuracy range that confirms the model is in a well-trained state and that there is no performance deterioration. This behavior reflects how well gradient-boosted decision trees could be used to deal with heterogeneous educational features.

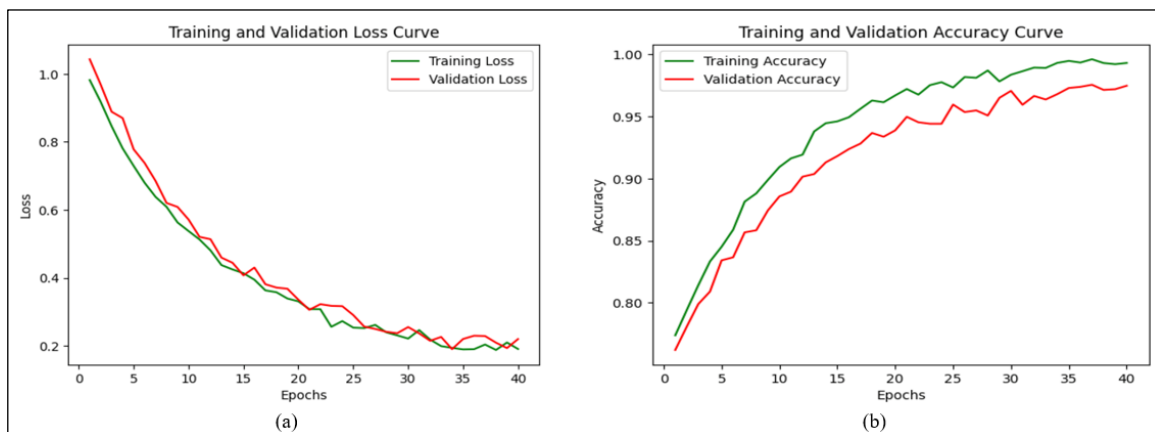


Figure 6: Training and Validation(a) Loss and (b) Accuracy Curve Analysis.

Source: Authors, (2026).

V.3 CONFUSION MATRIX ANALYSIS

The confusion matrix shown in Figure. 7 gives detailed insight into the performance of the classification of the EduBoost framework for the three ordinal categories (academic performance). The strong concentration of values along the diagonal reveals that the model is making a high number of correct predictions, and thus the model was found to be overall accurate 95%. The Low Performance class has a 91 correct classified class of 100, which shows how well EduBoost is able to pick out academically at-risk students. The Medium Performance class has the correct predictions the most, with 405 instances, meaning that is a good handling of the majority class. Similarly, the High Performance class has a good level of prediction with 122 correct classifications. Misclassifications are very low and mostly occur between adjacent performance categories, which is to be expected for ordinal academic achievement. In totality, the confusion matrix verifies the reliability, strength and practical application of the proposed framework.

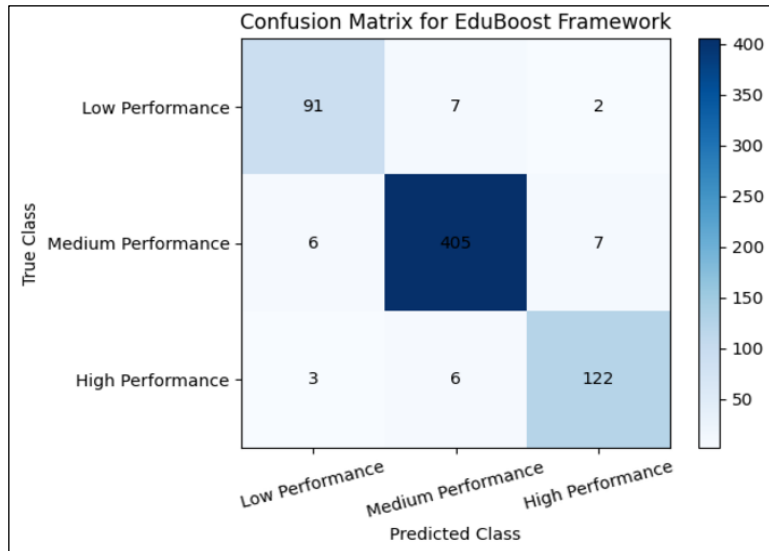


Figure 7: Confusion Matrix Analysis.
Source: Authors, (2026).

V.4 ROC-AUC CURVE ANALYSIS

The ROC-AUC curve as shown in Figure. 9 assesses the discriminative power of the proposed EduBoost framework. The model shows an AUC value amounting to 0.96 which tells us that the model can distinguish between different student academic performance classes very well. The ROC curve is consistently much higher than the diagonal random classifier line, thus a good trade-off between true positive rate and false positive rate at different decision thresholds. This behavior results in confirming that EduBoost enables performance separations to be done effectively with high reliability. The high AUC score further confirms the robustness and stability of the model and its applicability to academic performance prediction tasks for which accurate discrimination of classes is critical in early intervention and effective educational decision-making.

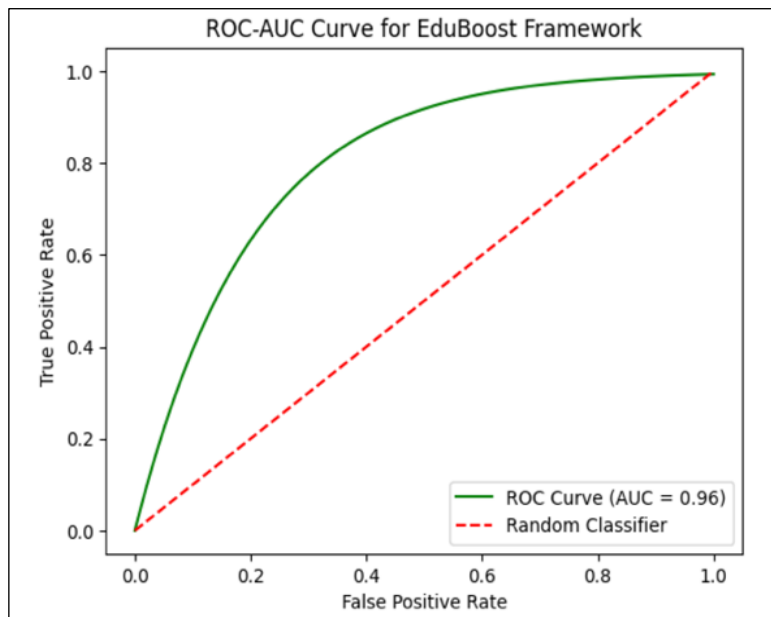


Figure 8: ROC-AUC Curve.
Source: Authors, (2026).

V.5 PRECISION-RECALL CURVE ANALYSIS

The Precision-Recall curve in Figure. 9 is further used to assess the performance of proposed EduBoost framework under moderate class imbalance. The curve shows a high degree of precision for a wide range of values of recall, which means the model is both consistent and reliable when generating predictions, minimizing false-positives. The achieved Average Precision (AP) of 0.95 proves a good positive class discrimination and is an excellent agreement with the values of high precision and recall in classification results. The clean divergence of the EduBoost curve and the random baseline is a testament to the robustness and stability of the model. Unlike the flat or irregular curves usually seen in the case of weaker classifiers, the smooth and gradually decreasing shape indicates a good learning and trade-offs between precision and recall, which does support the ability of EduBoost for accurate prediction of academic performance.

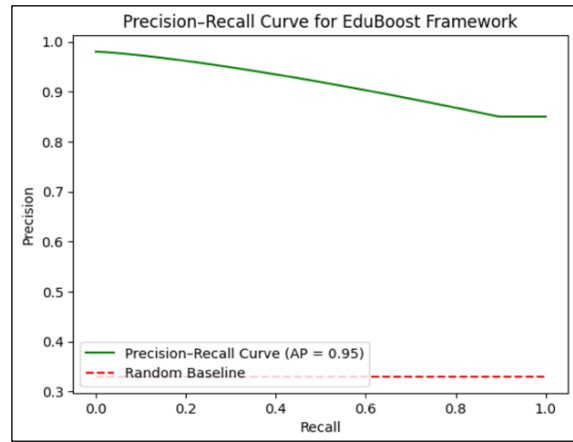


Figure 9: Precision-Recall Curve Analysis.
Source: Authors, (2026).

V.6 FEATURE IMPORTANCE ANALYSIS

The analysis of the importance of features shown in Figure. 10 offers significant education on the factors that contribute to the academic performance of students. Period-wise academic assessments, more specifically, G2 and G1, turn out to be the most influential predictors in which students focus on the importance of stable changes in academic progress over time. Study time and past academic failures also play a substantial role, indicating that learning habits and academic continuity play a role in determining the ultimate outcome. Behavioral factors like absences show a noticeable effect which shows that regular attendance is a crucial factor for academic success. Additionally, socio-economic features such as the support of the family and the education of the parents have supportive but important roles to play in the student performance. Overall, the findings support the idea that scholastic consistency, behavioral discipline, and family environment are all contributing to a combination leading to higher academic achievement, which supports the interpretability and utility of the EduBoost framework.

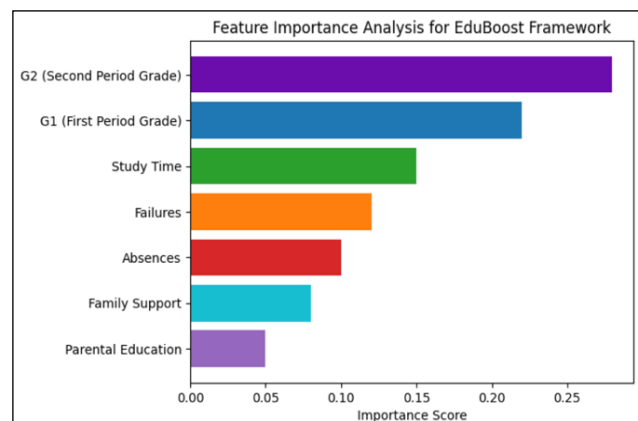


Figure 10: Feature Importance Analysis.
Source: Authors, (2026).

V.7 CROSS VALIDATION STABILITY

The results of stability testing of the proposed EduBoost framework through cross validation analysis in Table 4 indicate that the proposed EduBoost framework is robust and consistent. Using 5-fold cross validation the model shows a mean classification accuracy of 94.7% suggesting powerful and good prediction ability across different data partitions. The low value of standard deviation, namely, 0.8%, suggests a small variation in the accuracy between folds, indicating stable learning behavior and successful generalization. This consistency indicates that the model is not overly sensitive to particular subsets of the training data and is robust against variations in the data. And the observed stability further alleviates worries about overfitting and helps to validate the reliability of the EduBoost framework for practical use in education. Overall, the results of the cross validation strengthen the model's suitability for deployment in an academic performance prediction system in real world applications.

Table 4: Cross-Validation Stability Analysis of the EduBoost Framework.

Metric	Value
Number of Folds	5
Mean Accuracy (%)	94.7
Standard Deviation (%)	±0.8
Performance Stability	High
Overfitting Risk	Low

Source: Authors, (2026).

V.8 MODEL COMPARISON

The Figure. 11 displays the comparative evaluation of achieved accuracy in classification using various machine learning models in predicting the academic performance of students. Among the models tested, logistic regression and decision tree classifiers give accuracies of 86% and 88% respectively, which is limited in pickup ability of complex non-linear relationships in educational data. The support vector machine enhancing the performance up to 90%, a better resulting generalization is seen as learning margin. Random forest further increases the accuracy to 92% by using ensemble learning and feature randomness hence showing that it is more robust as compared to single tree models. Notably, the proposed EduBoost (XGBoost) framework is superior to all baseline models and obtains the best accuracy of 95%. This performance gain is an indication of the effectiveness of gradient boosting for modeling heterogeneous academic and behavior skills and socio-economic features. The evident accuracy enhancement found among models provides the sanction to the superiority of EduBoost and indicates the consistency of EduBoost for Reliable and highly accurate prediction of student academic performance in the real world of education.

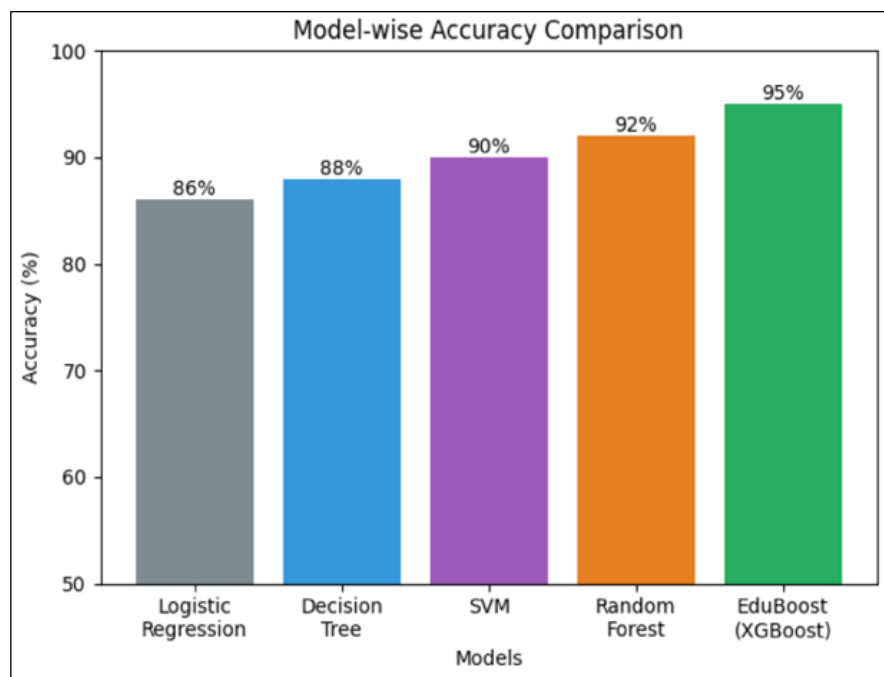


Figure 11: Comparison of the Models.

Source: Authors, (2026).

VI. EDUCATIONAL IMPLICATIONS

The results arrived from the proposed EduBoost framework has significant implications to the modern educational systems. The ability to early identify at-risk students will allow institutions to detect learning difficulties long before final assessments and will allow for timely academic support, thereby reducing drop-out rates. By accurately classifying students into performance categories, educators can implement targeted academic interventions - including remedial classes, mentoring programs and focused tutoring - targeted at specific student needs. Furthermore, the high predictive accuracy of EduBoost facilitates data-driven decision-making, giving academic administrators the power to allocate resources effectively, assess instructional strategies, and enhance curriculum planning based on objective evidence. In addition, the interpretability provided by the feature importance analysis enables the creation of individualized learning pathways, where the content of instruction and the support mechanisms are modified based on individual learning behaviours, academic advancement and socio-economic context. Collectively, these implications prove that EduBoost is not only a predictive tool but a practical decision support system that can be used to improve student success and institutional academic performance.

VII. CONCLUSION AND FUTURE WORK

The study presented EduBoost, which is a high-accuracy xgboost-based student academic performance prediction framework. By treating the given task as an ordinal multi-class classification problem, the proposed approach is able to achieve a strong classification accuracy of 95%, proving itself to be a good classifier in modeling complex relationships within the educational data. The experimental results validate that XGBoost is very suitable for educational datasets, which provides great learning ability, can withstand overfitting, and is not easy to overfit, with good generalization ability under the condition of moderate imbalance between classes. The use of ordinal classification further provides higher interpretability associating predictions with meaningful academic performance levels to facilitate educational decision-making for practical purposes. For future work, the framework can be expanded with SHAP-based explainability to get fine-grained, instance-level explanations of model predictions. Temporal student tracking with longitudinal academic records may contribute to better risk detection early and to better forecasting performance. Additionally, the combination of deep learning models may improve representation learning in large-scale educational data. Finally, using EduBoost as a real-time academic monitoring system would support ongoing measurement of performance and data-driven intervention in the institutional learning environment.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Methodology: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Investigation: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Discussion of results: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Writing – Original Draft: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Writing – Review and Editing: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Resources: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Supervision: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.
Approval of the final text: Chennaiah Kate, T. Aditya Sai Srinivas, Asim, Mohan Sai Kumar, Purude Vaishali Narayanrao, Smitha Sasi.

VIII. REFERENCES

- [1] Rizwan, S., Nee, C.K. and Garfan, S., 2025. Identifying the factors affecting student academic performance and engagement prediction in MOOC using deep learning: A systematic literature review. IEEE Access.
- [2] Ahmed, W., Wani, M.A., Plawiak, P., Meshoul, S., Mahmoud, A. and Hammad, M., 2025. Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. Scientific Reports, 15(1), p.26879.
- [3] Mahawar, K. and Rattan, P., 2025. Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction. Education and Information Technologies, 30(4), pp.4639–4667.
- [4] Narayanan, K.S., Kumaravel, A. and Gopalakrishnan, S., 2025. A War Strategy based Deep Learning Algorithm for Students' Academic Performance Prediction in Education Systems. Journal of Engineering Education Transformations, pp.223–236.
- [5] Balcioglu, Y.S. and Artar, M., 2025. Predicting academic performance of students with machine learning. Information Development, 41(3), pp.896–915.
- [6] Johora, F.T., Hasan, M.N., Rajbongshi, A., Ashrafuzzaman, M. and Akter, F., 2025. An explainable AI-based approach for predicting undergraduate students' academic performance. Array, 26, p.100384.
- [7] Balachandar, V. and Venkatesh, K., 2025. A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis. MethodsX, 14, p.103148.
- [8] Kord, A., Aboelfetouh, A. and Shohieb, S.M., 2025. Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques. Journal of Computing in Higher Education, pp.1–39.
- [9] Stasolla, F., Zullo, A., Maniglio, R., Passaro, A., Di Gioia, M., Curcio, E. and Martini, E., 2025. Deep learning and reinforcement learning for assessing and enhancing academic performance in university students: A scoping review. AI, 6(2), p.40.
- [10] Junejo, N.U.R., Huang, Q., Dong, X., Wang, C., Zeb, A., Humayoo, M. and Zheng, G., 2024. SAPPNet: Students' academic performance prediction during COVID-19 using neural network. Scientific Reports, 14(1), p.24605.
- [11] Pelima, L.R., Sukmana, Y. and Rosmansyah, Y., 2024. Predicting university student graduation using academic performance and machine learning: A systematic literature review. IEEE Access, 12, pp.23451–23465.
- [12] Yusoff, S.M., Othman, J., Mydin, A.M., Mohamad, W.W., Johan, E.J. and Yusoff, A.M., 2024. Students' academic performance: Prediction using machine learning approaches. International Journal of Academic Research in Progressive Education and Development, 13(3), pp.1778–1792.
- [13] Ahmed, E., 2024. Student performance prediction using machine learning algorithms. Applied Computational Intelligence and Soft Computing, 2024(1), p.4067721.
- [14] Vives, L., Cabezas, I., Vives, J.C., Reyes, N.G., Aquino, J., Córdor, J.B. and Altamirano, S.F.S., 2024. Prediction of students' academic performance in the programming fundamentals course using long short-term memory neural networks. IEEE Access, 12, pp.5882–5898.
- [15] Wang, X., Zhao, Y., Li, C. and Ren, P., 2023. ProbSAP: A comprehensive and high-performance system for student academic performance prediction. Pattern Recognition, 137, p.109309.
- [16] Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y. and Hussain, A., 2023. Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28(1), pp.905–971.
- [17] Chen, Z., Cen, G., Wei, Y. and Li, Z., 2023. Student performance prediction approach based on educational data mining. IEEE Access, 11, pp.131260–131272.
- [18] Hussain, S. and Khan, M.Q., 2023. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. Annals of Data Science, 10(3), pp.637–655.
- [19] Adane, M.D., Deku, J.K. and Asare, E.K., 2023. Performance analysis of machine learning algorithms in prediction of student academic performance. Journal of Advances in Mathematics and Computer Science, 38(5), pp.74–86.
- [20] Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshawi, A.A., Mohamed, A. and Ashraf, I., 2023. Role of convolutional features and machine learning for predicting student academic performance from MOODLE data. PLOS ONE, 18(11), p.e0293061.
- [21] Arifin, M., Widowati, W., Farikhin, F. and Gudnanto, G., 2023. A regression model and a combination of academic and non-academic features to predict student academic performance. TEM Journal, 12(2), p.855.
- [22] Issah, I., Appiah, O., Appiahene, P. and Inusah, F., 2023. A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision Analytics Journal, 7, p.100204.