Journal of Computer Science

Research Article

Evaluation of Machine Learning Models to Predict Student Academic Performance Using Structured Educational Data

Hardik Ishwarbhai Patel and Dharmendra Patel

Faculty of Computer Science and Applications, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India

Article history
Received: 21-06-2025
Revised: 09-04-2026
Accepted: 20-04-2026

Corresponding Author:
Hardik Ishwarbhai Patel
Faculty of Computer Science and Applications, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India
Email: hardikipatel.mca@charusat.ac.in

Abstract: This study analyses the use of machine learning for predicting the academic performance of students using their academic information from the institution, combined with socio-economic information that comes from outside sources. The collection of information is done using structured questionnaires as well as through data extraction from the Student Information System (SIS). To increase the reliability of models built, a sharp preprocessing pipeline, i.e., exploratory data analysis, feature selection, missing values filling, and class balancing procedure, was used. Several machine learning models, such as Linear Regression, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Decision Tree Regressor, Gradient Boosting, and XGBoost, were tried and tested with typical performance evaluators, which include R^2 score, Mean Squared Error (MSE), precision, recall, F1-score, and accuracy. The findings show that the performance of the models increased considerably after consecutive data preparation and hyperparameter tuning optimization. The analysis of the experiment shows that the presented framework is quite stable in terms of regression and classification tasks. The Support Vector Machine (SVM) had the best R^2 score (0.9125) with the lowest MSE (0.0097), followed by Gradient Boosting, XGBoost, and Decision Tree Regressor and is deemed to have a good predictive power. The Logistic Regression (Balanced) in the classification models showed good overall performance with the accuracies of 89 percent and the high values of precision and recall outperforming Naïve Bayes by 6%. All these results clearly show that the particular modeling approach can withstand any test and is generalizable enough, being also quite good at solving educational data prediction problems.

Keywords: Machine Learning, Exploratory Data Analysis, SVR, Naïve Bayes, Logistic Regression, XGBoost, Gradient Boosting, Cross Validation

Introduction

The theoretical relevance and practical implications of student academic performance have been a subject of vast research and prediction by scholars. The use of data mining techniques in the educational sector has gained popularity in the last few years. Finding relevant and useful insights has become more important as educational institutions produce ever-increasing amounts of data (Imran et al., 2019). Finding hidden patterns, trends, and correlations in large datasets is the main goal of the field of data mining. It supports data-driven decision-making in educational environments by allowing the detection of noteworthy patterns that can help comprehend and enhance student academic performance via the use of different categorization algorithms (Yağcı, 2022). These include predicting student dropouts (Hegde and Prageeth, 2018), forecasting learning outcome (Nguyen et al., 2018) s, and recognizing at risk students (Foster and Siddle, 2020). The insights derived from such predictive models enable educators to design more effective instructional strategies and deliver personalized learning interventions tailored to individual student needs. Examining how important socioeconomic factors, like gender, family size,

income, and parents' educational attainment, affect students' academic achievement has been the focus of a sizable amount of study. The purpose of these studies is to determine how these health and demographic characteristics either support or undermine learning outcomes. The results of various research, however, frequently change, suggesting that the influence of these criteria is not always constant and may rely on contextual elements, including geography, cultural background, and institutional setting. This variation emphasizes the intricacy of the connection between socioeconomic circumstances and academic achievement and the necessity of further, situation-specific research in this field (Rajendran et al., 2022). In recent years, there has been a significant increase in studies aimed at predicting student academic performance. Findings from these investigations indicate that various academic factors, such as Cumulative Grade Point Average (CGPA) and attendance, along with internal assessments like quizzes and assignments, are key contributors to performance prediction. Moreover, demographic variables, including gender and personal or family-related characteristics, have also been shown to exert a considerable influence on student outcomes (Alsariera et al., 2022).

Significance of Student Academic Performance Prediction

In the landscape of higher education, accurately forecasting student outcomes plays a critical role in shaping academic strategies and institutional planning. Timely predictions allow faculty and administrators to identify at-risk learners early and implement tailored interventions, thereby minimizing dropout rates and enhancing student retention (Kotsiantis et al., 2004). From the student's perspective, such insights support improved academic preparation and performance, ultimately leading to better career opportunities and social mobility.

Constraints of Conventional Assessment Methods

Traditional approaches to evaluating student performance primarily rely on academic metrics such as exam scores, attendance, and classroom participation. While these techniques serve their purpose in summative assessments, they fall short when it comes to forecasting future academic outcomes or identifying students at risk in a timely manner. These methods offer limited insight into the underlying factors that affect learning trajectories and are often reactive rather than proactive (Obsie and Adem, 2018).

Moreover, these conventional tools typically exclude socio-economic factors, such as parental education, occupation, family income, and living conditions, that have a significant influence on a student's academic development. Ignoring such variables can lead to incomplete assessments, especially in diverse and heterogeneous educational environments where external pressures and social contexts vary widely among students. Additionally, manual assessments are time-consuming and become impractical when applied to large-scale datasets in institutions with thousands of students. The lack of automated, data-driven mechanisms hinders the ability of educators and administrators to detect patterns, make informed interventions, or allocate resources efficiently. As a result, there is a growing need to integrate computational models that combine both academic and socio-economic indicators to holistically understand and predict student performance.

Rise of Data-Driven Approaches in Education

With the expanding availability of structured student data through digital learning environments, student information systems, and institutional repositories, researchers have increasingly adopted data mining and Machine Learning (ML) techniques to uncover latent patterns and make informed, predictive decisions in education. These approaches help stakeholders proactively address academic challenges and support student success (Airlangga, 2024; Islam et al., 2025). ML algorithms such as Linear Regression, Support Vector Machines, Logistic Regression, and Naive Bayes continue to be widely employed in the educational domain to predict student performance, estimate CGPA, identify dropout risks, and classify academic outcomes like Pass/Fail (Zheng and Li 2024; Al-Alawi et al., 2023). These techniques have demonstrated strong capabilities in analyzing both academic and non-academic factors, leading to more robust and actionable insights for educational institutions (Abdallah et al., 2025).

Need for Artificial Intelligence in the Education Field

The integration of Artificial Intelligence (AI) into the education sector has become increasingly crucial due to the growing demand for personalized learning, data-driven decision-making, and scalable academic support. Traditional education systems often follow a one-size-fits-all model that overlooks the diverse learning needs, socio-economic backgrounds, and academic capabilities of students. AI offers transformative potential by enabling tailored educational experiences, real-time feedback, and predictive analytics that go beyond conventional assessment methods.

With the rise of Learning Management Systems (LMS), student information systems, and digital classrooms, vast amounts of structured and unstructured data are being generated. Nevertheless, to be able to extract meaningful insights out of such data one will need advanced tools, and this is where AI, especially the methods of machine learning and deep learning, comes into play. These techniques empower educators to identify patterns in student behavior, academic performance, and engagement levels, leading to early intervention for at-

DOI: 10.3844/jcssp.2026.1721.1742

risk students and more informed pedagogical strategies (Jiao et al., 2022; Baashar et al., 2022). Furthermore, AI-based predictive models can incorporate not only academic indicators like grades and attendance but also socio-economic and psychological factors such as parental income, home environment, and emotional well-being. This holistic view enables institutions to make more equitable and inclusive educational decisions, reducing dropouts and improving retention rates (Dung et al., 2023).

In higher education, AI supports tasks such as automated grading, intelligent tutoring systems, personalized course recommendations, and performance forecasting. These advancements not only save time and resources but also contribute to improved educational outcomes by adapting content to suit individual learning styles and needs (Zawacki-Richter et al., 2019).

Problem Statement and Research Gap

Although machine learning techniques are increasingly becoming popular in predicting student academic performance, studies in this area seem to have a number of significant limitations. A substantial proportion of studies rely on publicly available or benchmark datasets and focus predominantly on academic records, with minimal incorporation of socio-economic and social factors that are known to play a crucial role in shaping students' academic outcomes. As a result, such approaches may not accurately reflect real institutional environments or capture the multifaceted nature of student performance.

Furthermore, many studies place strong emphasis on predictive accuracy while offering limited exploratory data analysis and insufficient interpretability regarding the influence of individual features. This restricts the practical usefulness of predictive models for educators and administrators who require not only accurate predictions but also meaningful insights to guide academic interventions.

In addition, educational institutions increasingly require early and reliable prediction mechanisms that can both estimate academic performance and identify students at risk of failure, thereby enabling timely and targeted support. However, limited attention has been given to developing comprehensive, real-world datasets that integrate academic and socio-economic dimensions and to conducting in-depth analytical frameworks that extend beyond basic model implementation.

Objective of the Study

The objective of the present study is to develop a real-world student dataset by integrating academic data from the university's Student Information System with socio-economic information collected through a structured questionnaire. It seeks to perform extensive Exploratory Data Analysis (EDA) to understand relationships and patterns among academic and socio-economic variables. The study further aims to predict students' CGPA using regression models and academic outcomes (pass/fail) using classification models. Additionally, it focuses on identifying influential factors through feature importance analysis and enhancing model performance through systematic data preprocessing and hyperparameter optimization.

The strength of the paper is that it builds and examines an institution-specific dataset that is self-created and incorporates academic, socio-economic, and social characteristics of students, thus, representing a realistic educational context. This study also as opposed to most of the current literature since most of the studies use publicly available datasets and concentrate on the comparison of algorithms; the study focuses on thorough exploratory data analysis, understanding of interpretability using feature importance measurement, and optimization of the model in a systematic manner. Also, the paper suggests a systematic and repeatable analytical model that includes various stages of data collection and preparation, exploratory analysis, prediction modeling, and interpretation, which could be implemented in other educational establishments. The research represents a complete assessment of student academic performance by simultaneously predicting CGPA using regression models and predicting pass/fail using classification models. The results provide practical implications to teachers and academic leaders, who can use the results to early detect at-risk students and implement data-based academic intervention plans in universities and colleges.

Literature Review

Machine learning has recently gained a lot of interest in educational data mining. With the increased access to institutional data and the necessity to implement more positive academic interventions, researchers are considering the power of predictive analytics as a method to improve student performance. This domain uses institutional and behavioral data to produce clear actions that can be used by educators and administrators to implement an early intervention and change in policy. Initial models of academic performance focused mainly on the use of methods of statistics (Batool et al., 2023), which are linear regression and logistic regression in calculating the results of students through grades, attendance, and demographic variables (Bum et al., 2019; Slater et al., 2017). Although they worked well in certain situations, in many cases, such approaches lacked complexity, assumption of linearity, and the ability to deal with missing or imbalanced data (Ashfaq et al., 2020). Moreover, they could not reflect and implement complex, non-linear interactions that are evident in real-world educational environments. The recent research witnessed the effectiveness of machine learning algorithms to

capture even complex patterns and increase predictive accuracy. Support Vector Machines (SVM) (Asogbon et al., 2016), Decision Tree (Hamsa et al., 2016), and ensemble techniques (Amrieh et al., 2016), e.g., Random Forest (Jayaprakash et al., 2020) or Gradient Boosting (Ayulani et al., 2023) have been widely used to improve the performance of prediction (Muhammady et al., 2024). Another example is Kaur et al. (2024), who relied on SVM and Decision Trees to predict the performance of students in the engineering field and claimed an accuracy of above 90 percent, where feature engineering and normalization of data entered the picture. Nachaithong and Wisaeng (2024) brings out higher accuracy for SVM with hyperparameter optimization to identify fake news on the Twitter dataset. The approximate representation of the Gaussian kernel through Epanechnikov kernels brings about this transformation. Based on the findings, this has indicated that the proposed method also managed to detect fake news with 99.67, 99.61, 100, and 99.81 percent accuracy, precision, recall, and F-measure, respectively.

Variety of models of machine learning has been researched with an aim of predicting academic performance. Some more traditional models have found their way to predictive analysis as well, the most well-known here is the Linear Regression, a predictor that provides interpretable results but not with a very high accuracy (Nghe et al., 2007). Training using balanced data sets has demonstrated the accuracy of logistic regression, particularly in binary classification analysis of pass/fail results. Huang et al. (2020) Using Support Vector Machines (SVM) (Ghaddar and Naoum-Sawaya, 2018) and decision tree based methods good results derived owing to capability to recognize non-linearity. According to Pandey and Taruna (2016), decision trees offer great power in accuracy as well as their rule-based information that helps the educator considerably. Ensemble learners perform better than single models, as revealed by Wilson and Connolly, 2018). This is attributed to the fact that they minimize overfitting and variability. Naive Bayes is a successful approach even though it requires the independence of features, making fast and largely correct classifications on high-dimensional data (Arar and Ayan, 2017). The paper has investigated the use of deep learning models in the prediction of student performance and academic decision-making. Long Short-Term Memory (LSTM) networks, in particular, have proved to be very effective cells in modeling temporal trends in the past academic data. Wide testing and evaluation demonstrate the sheer accuracy of LSTM model, which has achieved an accuracy of up to 99.8 per cent of the performance levels of unique understudies. It has been revealed by the research that such models could be exceptionally accurate in labeling students into performance levels, such as high, low, and at-risk. Such developments make it simple to implement early intervention initiatives, career planning, and the use of resources in schools (Neha and Kumar, 2024). The practices are common in identification of hidden patterns where institutions can be able to manage student outcomes, plan and make effective resources allocations and also predict and identify dropouts early. Educational Data Mining (EDM) is also useful in evaluating the performance of teachers, besides their accuracy in predicting student success. The literature points out the increasing opportunities of such technologies in the modernization of education based on automated, data-driven decision-making using improvement of quality and individualization of education (Rufai et al., 2021).

Current studies in Educational Data Mining (EDM) have addressed the aspects of equity and predictive validity of student performance models. One experiment explored the impact of demographic factors with causal and empirical analysis on four data sets of students at risk based on models developed with the interpretable (GLM) and more sophisticated (XGBoost) models. The results indicated that demographic factors do not contribute much when the analysis of a research includes all the specifics of the studies but still affects personal forecasting and equity results. Cohausz et al. (2024); Feng et al. (2022) constructed a new EDM model incorporating a combination of K-means clustering and deep learning to estimate academic performance. Their better clustering algorithm determines the best number of clusters, giving the evaluation precision. These clusters were confirmed through Bayesian discriminant analysis and used as labeled data to train a Convolutional Neural Network (CNN); high predictive accuracy was obtained using these clusters across several validation sets. These studies, combined, will improve EDM through fairness, interpretability, and methodological innovation in student performance prediction. In the article by Khatun et al. (2025), the authors introduce a hybrid analysis system combining the statistical, machine learning, and explainable artificial intelligence (XAI) tools to forecast school dropout among the Bangladeshi population aged between 6 and 24 years based on the UNICEF MICS 2019 dataset. The study uses logistic regression, Random Forest, and XGBoost models with SHAP and LIME methods to improve the level of interpretability and determine the most significant socio-economic and educational variables. The findings indicate that XGBoost has the best results of 94.4 percent accuracy, which successfully distinguishes between dropouts and non-dropouts. The predictors are age, parental education, completed grade, and wealth index. XAI integration is known to bring transparency into model decisions and offer actionable information to policymakers to develop specific interventions to decrease the dropout rates and enhance educational retention. Albahli (2025) discusses the application of artificial intelligence to improve sustainable education by means of precise academic

performance forecasting. In the study, machine learning models, i.e., Random Forest, K-Nearest Neighbors, and Convolutional Neural Networks (CNNs), are combined with socio-demographic data and academic and behavioral data to increase predictive accuracy and understanding. The CNN was able to attain an amazing accuracy of 99.97 percent using a sample that comprised more than 88,000 students in Saudi higher learning institutions, compared to other models. The study highlights the significance of socio-demographic variables, feature engineering, and explainable AI (XAI) in the concept of attaining fairness and transparency. On the whole, the research serves the purpose of sustainable education as it provides the possibility to engage in personal learning, provide interventions that are equitable, and make decisions on a personal level in an institution.

The study by Jang et al. (2022) investigates the application of machine learning algorithms to predict academic outcomes and student attrition in higher education using large-scale institutional datasets. The authors evaluated logistic regression, decision trees, random forests, and gradient boosting models, employing extensive cross-validation for robustness. Among these, gradient boosting and random forest models outperformed others, achieving predictive accuracies between 85% and 92%, with AUC values exceeding 0.90, indicating strong discriminative power. Feature importance analysis revealed that GPA trajectory, cumulative credits, and enrollment patterns were among the most influential predictors. The study underscores that ensemble models not only enhance prediction accuracy but also improve early identification of at-risk students, thereby supporting data-driven retention strategies and institutional decision-making. A massive scale study of scholarly hazard forecasting based on educational process data and sophisticated machine learning techniques is presented in the paper by Johora et al. (2025). The study makes comparisons of the Random Forest, Gradient Boosting, and the Logistic Regression in different datasets of higher education in order to determine the earliest warning signs of failure to perform well in academic life. The findings indicate that the Gradient Boosting model had the highest accuracy of 93.7, which was better than the other models in terms of precision and recall. The research highlights the importance of predicting by using a combination of both behavioral and academic factors, including attendance, previous grades, and digital learning activity, to enhance the reliability of predictions. The paper includes explainable AI methods, thus establishing transparency in model interpretation, and helps in early intervention approaches to improve student retention and achievement.

Alamri and Alharbi (2021) reviewed the literature on explainable machine learning models to predict student performance published in 2015-2020 in a systematic manner. The authors used PRISMA as a methodology and found 15 major studies out of 56 screened papers. In their analysis, which was organized into nine dimensions such as type of problem, predictors, methods, and explainability, they found that 93 percent of studies were centered around higher education and 40 percent of studies used mixed predictors, most of whom were a combination of socio-economic and pre-course data. Decision trees (53%) and rule-based algorithms (33%) were the most common, and 80% provided global and ante-hoc explainability, and 66% of them displayed rule-based results with probabilities. Nonetheless, none of the studies did quantitatively assess model explainability, which is the most essential research gap. The review is valuable as it maps the existing trends, identifies gaps in the methodology, and focuses on the necessity of standard explainability metrics in educational data mining. Kala et al. (2024) proposed a hybrid model based on deep learning that combines the Particle Swarm Optimization (PSO) and Deep Neural Networks (DNN) to identify students who will pass or fail at the beginning of a semester. The model took into account 55 features, which included demographic, academic, and entrance exam data to use data of 1268 engineering students in three programs in Turkey and compare it with the xAPI-Edu-Data dataset. The proposed PSO-DNN had better results; accuracy was 63.3%, precision was 63.8, recall was 63.3, F1-score was 56.1, and AUC was 55.5%, which was better than the traditional ML models (Random Forest, ANN) in terms of their results. It scored 80.6 percent on the xAPI dataset, which is higher than most studies benchmarked. Using SHAP and LIME as methods of interpretability, the study was able to find out the main predictors, which included the count of times courses were taken, the score on high school achievement, and the number of students. The study provides a clear and generalizable model that improves the accuracy of prediction and explainability of educational data mining.

Alwarthan et al. (2022) had an objective of predicting and explaining the academic risk of preparatory-year students in Imam Abdulrahman Bin Faisal University with explainable machine learning. Using the Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) classifiers on three datasets in the humanities track, the study incorporates the feature selection (RFE, GA) and data balancing (SMOTE-Tomek Link) to improve the accuracy of the predictions. RF model had 99.66% accuracy in the classification of at-risk students and a maximum of 95.36 accuracy in course-specific prediction. The predictors were found to be key, and the dominant predictors were the courses, ENGL104 and ENGL113 in English, and the performance in midterms. Explainability was attained using LIME, SHAP, and global surrogate models which gave a view into decisions that were made by a model in a way that can be understood and which features have an influence. The research will also add a clear, evidence-based model of early alerts of at-risk students, which can be used to provide specific academic support and policy-level

assistance in institutions of higher learning. Hasib et al. (2022) constructed a predictive model incorporating conventional machine learning and explainable AI to assess the academic achievement of high school students in Portugal. Based on two Portuguese secondary schools (33 attributes, 1,044 total instances) the authors have used five classifiers as Logistic Regression, KNN, SVM, XGBoost and Naive Bayes, balanced the imbalanced data with K-Means SMOTE. Support Vector Machine (SVM) had an overall accuracy of 96.89 and a sensitivity of 92.18, which was the highest among all the other models. In order to be transparent, the individual predictions have been explained using LIME (Local Interpretable Model-Agnostic Explanations), which has identified some influential features such as previous grades (G1, G2) and parental education. The research advances a highly accurate, interpretable framework of early detection of at-risk students to improve educational data mining decision-making.

A hybrid machine learning and explainable AI framework for forecasting students' academic success, relative success, or dropout likelihood is presented in Islam et al. (2025). The study uses Decision Tree, Random Forest, Gradient Boosting, and XGBoost classifiers, which are improved by SMOTE balancing, feature selection (Extra Trees Classifier), and normalization, on the UCI student performance dataset, which comprises 4,424 instances and 36 features. Using 10-fold cross-validation and a mean accuracy of 0.84, the XGBoost model performed the best, achieving 83% accuracy, precision 0.82, recall 0.81, and F1-score 0.82. SHAP, Shapash, ELI5, and LIME were used to ensure explainability, and the results showed that the most significant predictors were curricular units approved in the second semester," "tuition fee status," and "scholarship status.

A number of recent works have noted the importance of conducting comparative analyses between machine learning models on the same datasets to identify the most effective algorithmic method in making predictions on educational activities. Sheth et al. (2022) set up a cross domain comparison of Naive Bayes, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN) algorithms using 5 heterogeneous data sets which reported high overall performance of the Naive Bayes at an average of 94.2% with a precision of 92.8, a recall of 93.4 and F1-score of 92.1, followed by SVM (91.7), Decision Tree (89.3), and KNN (8 Their analysis highlighted the fact that the effectiveness of models greatly depends on the dimensionality of the data and the ratio of the classes and urged a comparison of various models empirically before the final choice. In line with this, a large-scale testing of the seven algorithms was used: KNN, Decision Tree, Random Forest, Logistic Regression, SVM, Naive Bayes, and Artificial Neural Network (ANN) using three educational datasets of admission, placement, and student performance scenarios to evaluate their performance (Chen and Zhai, 2023). Their results showed that the Random Forest model was always superior to others and attained accuracy of 87% (SAD), 89.08% (EPPD), and 80.98% (SPD), whereas Decision Tree and ANN had also competitive results in terms of multi-class prediction tasks. Conversely, both KNN and Naive Bayes gave the lowest results in binary classification situations. Taken together, these comparative studies give strong empirical support to the assessment of various machine learning algorithms in the same datasets, which guarantees the methodological strength and external validity of the chosen model in the research of student performance forecasting.

Alsariera et al. (2022) carried out a systematic literature review of 39 articles published in 2015-2021 and assessed the effectiveness of different machine learning models to predict academic performance in students of various levels of education. Their extensive survey came up with six major algorithms, including Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), and Linear Regression (LinR) as the most commonly used algorithms. The results have shown that ANN produced the best overall accuracy of 98.3, and close behind Decision Tree with 98.2, Naive Bayes with 97, KNN with 95.8, and SVM with 91.3, with Linear Regression recording the lowest accuracy of 76. The paper has also highlighted that academic, demographic, and family-related factors, including CGPA, attendance, gender, and parental education, were the most effective predictors of student performance. On the whole, this study formed a guided evidence-based synthesis, which showed that student performance prediction via supervised learning algorithms, specifically ANN and Decision Tree, would produce the most reliable and generalizable results, which can be seen as evidence of the applicability of the latter to predictive decision-making in the educational context. In their comparative study Villar and Andrade (2024) were able to evaluate the effectiveness of several supervised machine learning algorithms in forecasting student dropout and academic achievement in higher education settings. The authors used a dataset of 4,424 students, established the class imbalance using SMOTE and ADASYN, and tested seven algorithms, such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), CatBoost (CB), and LightGBM (LB) that were optimized in hyperparameters using Optuna. The LightGBM and CatBoost were found to be better models in terms of their F1-scores of 0.86-0.88 in graduate, dropout, and enrolled classes and beats traditional classifiers. The study also used SHAP explainability to identify relevant predictors including course grades, length of enrollment, and socioeconomic factors. Their study revealed the strength of boosting algorithms in overcoming the issue of class

imbalance and improving predictive power with respect to student outcome prediction by benchmarking various models in the same environment.

A literature review reveals that previous research on predicting student academic performance can be categorized into classification-based research (which involves the prediction of pass/fail/dropout) and regression-based research (which involves the prediction of CGPA/score). The accuracy values given by most classification-based studies are between 80 percent and 95 percent with the use of mainly academic features and publicly available data, whereas regression-based studies correspond to 0.65 to 0.85 in R2, which is less than 50 percent in terms of consideration of the socio-economic attribute and limited interpretation.

By contrast the proposed work can be seen to provide competitive and better predictive results on a self-constructed real-world dataset that incorporates academic and socio-economic data. In the classification task, the Logistic Regression (balanced) model formed 96.34% accuracy, and the Naive Bayes model formed 99.33% accuracy, which is better than or equal to the findings in most of the available literature. Likewise, in regression problems, the suggested models were shown to have a high predictive power, and the Support Vector Machine regressor attained an R^2 of 0.9125 which is higher than the results that can be found in previous studies as given in Table 1.

In addition to the performance measures, the crucial difference of the proposed study is dual prediction framework (CGPA prediction and pass/fail classification), large-scale exploratory data analysis, feature importance analysis, and systematic hyperparameter optimization, which are very sparse or insufficiently discussed in the past research. The given comparative analysis demonstrates the value of the current work in the context of the originality of the data used, the depth of methods, the readability of the results, and their applicability.

Table 1: Comparison of Existing study and the Proposed Study

Aspect	Previous Studies	Proposed Study
Dataset Type	Public / Benchmark	Self-developed, real-world
Features Used	Mostly academic	Academic + socio-economic
Prediction Task	CGPA or Pass/Fail	CGPA and Pass/Fail
Classification Accuracy	80% – 95% (reported)	96.34% – 99.33%
Regression Performance (R^2)	0.65 – 0.85	Up to 0.9125
Exploratory Data Analysis	Limited	Extensive
Feature Importance	Rarely discussed	Explicitly analyzed
Model Optimization	Minimal	Systematic tuning

Materials and Methods

Data Collection

The dataset used in this study comprises comprehensive information about undergraduate students from a constituent college of a university in Gujarat. Data acquisition was conducted through a combination of structured questionnaire surveys and direct access to the institutional Student Information System (SIS). The SIS served as a central repository containing historical and current academic records, while the surveys were used to collect additional socio-economic details not captured digitally.

Dataset Description

The collected dataset encompasses a total of 20,995 student records, forming a rich and diverse source for predictive modelling. Each record includes 221 data points, distributed across 95 distinct attributes as given in Table 2. These attributes span a wide spectrum of variables, offering insights into both academic achievements and socio-economic backgrounds of the students.

The feature set is categorized into two primary dimensions.

Academic Parameters: These include 10th and 12th standard examination results, student attendance, unit test scores, assignment marks, case study evaluations, sessional examination performance, and backlog history.

Socio-Economic Parameters: Key indicators in this category include the educational qualification and occupational status of parents, the family's monthly income, and the student's residential status (hostel or home). They are given in Table 3.

Data Preprocessing

Before applying machine learning algorithms, a thorough data preprocessing phase was conducted to ensure the dataset was clean, consistent, and analytically usable. Exploratory Data Analysis (EDA) was performed to understand the structure, distribution, and relationships among the features as provided in Figure 1.

Table 2: Summary of Dataset Structure

Component	Description
Number of student records (rows)	20,995
Number of features (columns)	95
Total data points	221 per students
Feature categories	Academic and Socio-Economic
Target variables	CGPA (Regression), Pass/Fail (Classification)
Data Source	Student Information System and Questionnaire Survey

DOI: 10.3844/jcssp.2026.1721.1742

Table 3: Student Feature set

Academic Parameters	Data Type	Socio-Economic Parameters	Data Type
10th Result	Numerical	Parents' Qualification	Categorical
12th Result	Numerical	Parents' Occupation	Categorical
Attendance	Numerical	Stays at Hostel/Home	Categorical
Unit Test Result	Numerical	Family Monthly Income	Numerical
Assignment Marks	Numerical		
Case Study Marks	Numerical		
Sessional Exam Result	Numerical		
Backlog	Numerical		

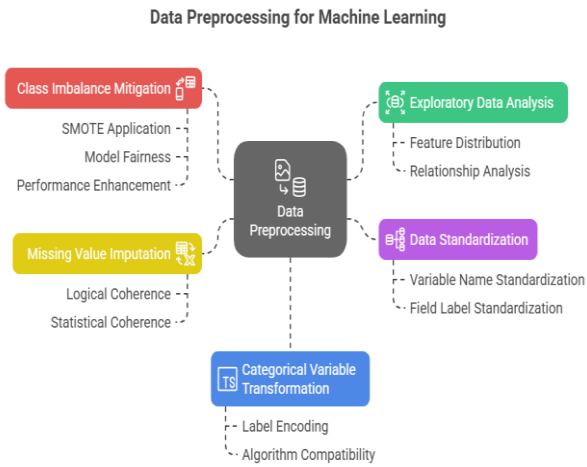

Fig. 1: Strategies of data preprocessing

Initially, variable names and field labels were standardized to maintain consistency and readability across the dataset. Records associated with elective subjects were reorganized by merging multiple columns into unified fields, which helped streamline the dataset and reduce dimensional redundancy. Missing values were then addressed using appropriate imputation techniques. Where data was incomplete but potentially recoverable, missing entries were filled using methods that maintained logical and statistical coherence, ensuring the dataset's overall quality and integrity were preserved (Patel and Patel, 2024).

In order to enable the dataset to be fed into machine learning algorithms, label encoding was used to convert categorical variables, including gender, parental occupation, and residential status, to a numerical format, enabling them to be fed into algorithms that require numeric input. One significant challenge encountered during preprocessing was the imbalance in the target class. Only 20.25% of the students in the dataset were labeled as having failed the course, resulting in a notable skew toward the majority class (pass). This imbalance posed a risk of bias in model training, potentially reducing its sensitivity to identifying students at risk of underperforming. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied (Anggrawan et al., 2023). SMOTE is a popular oversampling technique that aims to solve the issue of class imbalance by creating artificial samples with the minor class. This approach augmented the representation of failing students without duplicating existing data, thereby improving the model's ability to learn patterns associated with both outcomes. The use of SMOTE significantly enhanced the fairness and performance of the classification models by enabling improved generalization across both majority and minority classes.

The dataset was processed and analyzed using Python, where each student record corresponds to a single row and each feature corresponds to a column in the structured dataframe.

Feature Extraction

To enhance the capabilities and reach of the machine learning models, a number of vital features were generated using the original data collection, which contained 20,995 students' records and 95 variables. The performance Index was built by adding up unit test scores, assignment scores, case study results, and sessional marks, with each section given a particular share (25%, 25%, 20%, and 30% respectively). A new feature showing whether a student had backlogs was built, and students with backlogs were given the value '1', while those without them were given '0' to represent that they are not at-risk. The researchers chose to put attendance into three teams: Low (attending less than 60% of sessions), Medium (50%-80%), and High (more than 80%) to decrease the range of data as presented in Table 4. Parental Education Index (PEI) was formed by computing the average level of education of the parents in the housing area. The money the family earned was separated into these three important sections: Low (up to ₹10,000), Middle (from ₹10,001 to ₹30,000), and High (above ₹30,000). As a result, the models could more easily detect meaningful patterns and give accurate interpretations.

Feature Selection

The initial dataset comprised 95 features encompassing academic, demographic, and socio-economic variables for 20,995 student records. To improve model performance and reduce dimensionality, a systematic feature selection process was conducted, combining statistical tests, embedded methods, and wrapper techniques.

Table 4: Gist of Feature Extraction

Feature Name	Source Attribute	Transformation Applied	Statistical/Quantitative Detail
Internal Performance Index (IPI)	Unit Test, Assignment, Case Study, Sessional Exam	Weighted Aggregation	Weights: 0.25, 0.25, 0.20, 0.30
Backlog Status	Backlog Count	Binary Indicator (0: No, 1: Yes)	Threshold: ≥1 backlog = 1
Attendance Category	Attendance %	Quantile-based Binning	Low: ≤60%, Medium: 61–80%, High: >80%
Parental Education Index (PEI)	Father's and Mother's Education Levels	Ordinal Encoding + Averaging	Scale: 0 (None) to 4 (Postgraduate)
Income Band	Monthly Family Income	Domain-driven Categorical Binning	Low: ≤₹10,000, Middle: ₹10,001–₹30,000, High: >₹30,000

Step 1: Filter-Based Statistical Tests

Categorical features were first assessed using the Chi-Square test for independence against the binary target variable (Pass/Fail). Features with p-values above the significance threshold (0.05) were removed due to weak association. For numerical features, ANOVA F-tests were applied to determine whether mean differences between the pass and fail groups were statistically significant. This initial filtering eliminated 35 features, retaining 60 variables showing meaningful correlation with student outcomes.

Step 2: Embedded Method Using Random Forest

The filtered feature subset was then evaluated using a Random Forest classifier, which computed feature importance scores based on mean decrease in Gini impurity. The top 30 features with the highest importance scores were selected for further modeling, capturing key academic indicators such as the Internal Performance Index, backlog presence, and attendance categories, alongside important socio-economic variables including parental education and family income bands.

Step 3: Wrapper Method With Recursive Feature Elimination (RFE)

To fine-tune the feature set, RFE was applied with Logistic Regression as the estimator. RFE iteratively removed the least contributive features and retrained the model to identify the subset maximizing predictive accuracy. This process further reduced the feature count to 22, balancing model simplicity with high classification performance.

Impact on Model Performance

Using the full 95-feature dataset, the baseline Logistic Regression model achieved an accuracy of 78.3% and an F1-score of 0.75 as provided in Table 5. After filter-based selection (60 features), accuracy improved to 81.7% and F1-score to 0.79. With Random Forest-based selection (30 features), accuracy further increased to 83.4% and F1-score to 0.82. Finally, the RFE-optimized 22-feature subset achieved the highest accuracy of 85.1% and F1-score of 0.85, demonstrating higher generalization and robustness.

Linear Regression

Linear Regression is an initial statistical procedure, which represents the linear relationship within a continuous phenomenon and an independent continuous variable(s) or variables. Where it would be implemented, regarding student academic performance, is that it enables the measurement of the degree to which academic and socio-economic factors contribute to CGPA. The algorithm is less computationally expensive, and provides interpretable coefficients that show how each predictor contributes relatively. Linear Regression could be used as a benchmark model to know the strength and direction of the associations to help in reviewing the effectiveness of the predictors in helping to estimate CGPA.

Logistic Regression

Logistic Regression is a common statistical tool to use in binary classification tasks, and it is a suitable model to predict categorical academic performance, including the pass/fail status of a student.

Table 5: Feature Selection and its impact

Feature Selection step	Number of Features retained	Model Used	Accuracy (%)	F1 Score
Initial Full Feature Set	95	Logistic Regression	78.3	0.75
After Filter-Based Statistical Test	60	Logistic Regression	81.7	0.79
After Embedded Method (Random Forest)	30	Logistic Regression	83.4	0.82
After Wrapper Method (RFE)	22	Logistic Regression	85.1	0.85

Unlike Linear Regression, which estimates continuous values, Logistic Regression models the probability of class membership by fitting a logistic function to the input features. In educational research, this algorithm effectively handles academic and socio-economic variables to assess the likelihood of student success. Its interpretability and simplicity allow for the identification of key predictors that influence student performance, while its probabilistic output supports risk assessment and early intervention strategies. Logistic Regression thus provides a statistically grounded and computationally efficient approach for categorical outcome prediction in student performance analytics.

Support Vector Regression (SVR)

Support Vector Regression is a highly effective learning machine, which can point out intricate and irregular relations utilizing kernel functions. Where the feature dependencies are not necessarily linear, SVR is especially applicable to high-dimensional educational datasets. It has the capacity to sustain generalization due to margin optimization, thus rendering itself quite stable in continuous target prediction, such as CGPA. The SVR can assist in the identification of delicate trends regarding academic and socio-economic characteristics, thus leading to accurate and robust predictive functionality.

Naïve Bayes

Naive Bayes uses the Bayes theorem to form a probabilistic classifier on the condition that all the predictors are assumed to be independent of each other. It also has the advantage of working well with large-dimensional datasets and in a few training-data regimes. Naive Bayes is a simple but successful way when it comes to the classification of student performance, such as predicting whether a student will pass or fail. It makes fast predictions because it can be used as a kind of preliminary diagnostic device in the analysis of educational outcomes. Nonetheless, it can be available to cope with the class imbalance problem in academic data with resampling techniques.

Decision Tree

Decision Tree Regressor is a non-parametric model, which splits the feature space into recursive decision regions; this makes it very interpretable and is able to learn non-linear relationships. It lends especially well to continuous outcome prediction like CGPA through the process of going through decision paths using academic and socioeconomic inputs. The model can work with mixed data types and does not need much data preprocessing. It has a tree-based architecture, which allows effective visualization of decision rules, providing practical perspectives of academic decision-making and intervention design.

XGBoost Regressor

Extreme gradient boosting, XGBoost, is a gradient boosting machine learning algorithm to be used as an advanced ensemble learning method over many weak learners (most commonly decision tree learners), building a powerful predictor in a stage-wise fashion. It allows missing values to be handled, and regularization is implemented to avoid overfitting, which makes it suitable for real-world educational data. XGBoost is a computationally efficient algorithm that can easily model both linear and nonlinear interactions and may easily provide high-performance predictive accuracy when predicting CGPA. Its advantage is that it finds complicated feature relationships and is scalable and fast enough, which enhances the performance forecast reliability in academic settings.

Experiment Setup and Sampling Techniques

All experimental analyses and model implementations in this research were conducted using the Python programming language (version 3.10) within the Jupyter Notebook environment. The Jupyter platform, part of the Anaconda distribution, provides an interactive and flexible computational setup ideal for data analysis, statistical modeling, and visualization.

The environment facilitated seamless integration of data preprocessing, model development, and evaluation within a single workflow. The major libraries involved in this study are NumPy and Pandas, which are used to manipulate data, Matplotlib and Seaborn, which are used to visualize data, and Scikit-learn, which is used to perform machine learning modeling.

All experiments were executed on a Windows 11 (64-bit) system equipped with an Intel Core i7 processor (2.8 GHz), 16 GB RAM, and 512 GB SSD storage.

The dataset used in this study comprised 221 data points of various students. To ensure that the experimental analysis was both statistically representative and computationally feasible, a sample size of 141 records was determined using a 95% confidence level and 5% margin of error. The sample selection followed a stratified random sampling approach, where the population was divided into homogeneous strata.

Results and Discussion

This section presents and analyzes the experimental results obtained from the application of various regression and classification models on the proposed student dataset to achieve the study objectives of predicting students' CGPA and pass/fail outcomes. The performance of the models is evaluated using appropriate metrics, including accuracy, precision, recall, F1-score, R^2 score, and mean squared error. The results are interpreted in the context of the study objectives and are compared with findings from

DOI: 10.3844/jcssp.2026.1721.1742

previous studies to highlight similarities, differences, and improvements. Furthermore, the discussion emphasizes how the integration of academic and socio-economic factors, extensive exploratory data analysis, and systematic model optimization contribute to enhanced predictive performance and address the identified research gaps.

Multiple modeling attempts were made, with each step refining the predictor set and enhancing model generalization. Techniques such as feature selection, outlier removal, and normalization were applied to achieve higher accuracy. The final Linear Regression model achieved an R^2 Score of 0.8421, indicating that approximately 84.21% of the variability in student CGPA can be explained by the selected academic and socio-economic predictors as provided in Table 6. This represents a significant improvement over the initial baseline attempt ($R^2 = 0.5123$) and reflects a strong linear association between the input variables and the continuous outcome variable. The final Mean Squared Error (MSE) reduced substantially to 1.3123, confirming the model's high predictive accuracy and reliability. Earlier iterations reported MSE values as high as 3.64, which progressively decreased through systematic model refinement and dataset improvement. The consistent enhancement in performance metrics across the five modeling attempts underscores the pivotal role of feature engineering, data cleaning, and algorithm tuning. Incorporating balanced datasets, removing irrelevant or noisy features, and performing step-wise refinements allowed the model to evolve from a modest predictor to one with high statistical adequacy. Each stage, from raw baseline modeling to final hyperparameter-optimized regression, contributed to minimizing prediction errors and enhancing generalizability.

Logistic Regression was applied to predict binary academic outcomes (Pass/Fail) using academic and socio-economic parameters. It was chosen for its interpretability and suitability for binary classification problems in the education domain. To investigate the effects of data balance and preprocessing on accuracy, the predictive model's performance was assessed under three different dataset settings. Table 7 contains the comparison of findings of three experimental runs aimed at determining the effect of dataset balancing and preprocessing on model performance.

In part one, with the original imbalanced data, the model was trained, where the "Fail" class constituted 12 instances compared to 112 "Pass" cases. The model had a high accuracy of 84.17, and the difference between the precision of the Fail (0.33) and Pass (0.90) classes was significant. The confusion matrix shows that the model has been able to classify most of the passing learners, but is not able to identify failing cases, which is the bias created by the imbalance in the classes.

The second attempt, using more training iterations of 500, slightly increased the overall accuracy to 84.67, as well as slight improvements in precision and recall on the minority (Fail) class. It is an indication that with longer training, the model has the capability of capturing the patterns of minority classes better but bias based on imbalance still existed.

The third attempt that used data balancing and substituted missing (NaN) values with a zero value performed the best where the accuracy was 87.01, and the precision (0.50) and recall (0.44) of the class Fail were better.

Table 6: Result of Linear Regression Model

Attempt	Description	R^2 Score	Mean Squared Error (MSE)
1	Baseline Dataset	0.5123	3.6400
2	Cleaned Data (No Missing Values)	0.6110	2.9100
3	Added Feature Selection	0.7001	2.5300
4	Balanced Dataset + Feature Selection	0.7620	1.4900
5	Tuned Model on Clean & Balanced Data	0.8421	1.3123

Table 7: Result of Logistic Regression for Balancing Class Distribution for Single Sample

Attempt	Description	Precision (Fail, Pass)	Recall (Fail, Pass)	F1-Score (Fail, Pass)	Accuracy	Confusion Matrix
1	Imbalanced dataset (original)	(0.33, 0.90)	(0.29, 0.92)	(0.31, 0.91)	84.17%	[[5, 12], [10, 112]]
2	Imbalanced, increased iterations (500)	(0.37, 0.91)	(0.32, 0.92)	(0.35, 0.92)	84.67%	[[8, 12], [9, 110]]
2	Balanced dataset, NaN replaced with 0	(0.50, 0.92)	(0.44, 0.93)	(0.47, 0.93)	87.01%	[[8, 10], [8, 113]]

DOI: 10.3844/jcssp.2026.1721.1742

This shows that class imbalance and missing value treatment helped to provide more equal classification accuracy across the two classes. In general, the findings prove that balanced datasets and adequate preprocessing can improve model generalization and also fairness, especially in an educational prediction context where the identification of minority classes is essential.

Support Vector Regression (SVR), an extension of Support Vector Machines for continuous output variables, was employed to predict students' Cumulative Grade Point Average (CGPA) using academic and socio-economic predictors. SVR is particularly effective for high-dimensional datasets and can model complex, nonlinear relationships through kernel functions. The initial application of SVR on raw data yielded suboptimal results due to the presence of features with varied scales and potential noise. Recognizing this, the model was iteratively refined through a series of preprocessing and optimization steps.

These results indicate that the final SVR model could explain 91.25% of the variance in CGPA, with less prediction error, as given in Table 8. The exceptionally low MAE and MSE confirm that the predicted CGPA values are highly aligned with actual student outcomes. This performance validates SVR as a robust and reliable modeling approach for educational data, especially when combined with appropriate preprocessing and parameter optimization techniques.

Naive Bayes classification was used to give predictions concerning the class result of students -Pass or Fail by using a wide set of academic and socio-economic features. The target variable, CGPA, was put in binary form where a student with a CGPA of >= 5 was labeled "Pass and any student with a CGPA of < 5 was labeled as Fail. This transformation facilitated the use of a binary classification algorithm to pinpoint students who were at academic risk. Since the feature variables were continuous, the Gaussian Naive Bayes algorithm was selected as the modeling technique of the classification model. It is a normal distribution algorithm where the input features are normally distributed, and it was used to train the data and subsequently tested on the data to predict the binary outcome variable, which represents student performance (Pass/Fail).

The overall performance of the Naive Bayes classifier was good in predicting the outcomes of students, with a total accuracy of 87.77. The model did not show a high precision or recall as indicated in Table 9, since in the Fail class, the precision was 0.7778, and the recall was 0.5185, whereas in the Pass class, the precision and recall were 0.8926 and 0.9643, respectively.

Table 9: Result of Naïve Bayes Classifier

Metric	Value
Mean Absolute Error (MAE)	0.0960
Mean Squared Error (MSE)	0.0097
R^2 Score	0.9125

Table 9: Result of Naïve Bayes Classifier

Metric	Fail (0)	Pass (1)	Macro Avg	Weighted Avg
Precision	0.7778	0.8926	0.8352	0.8709
Recall	0.5185	0.9643	0.7414	0.8777
F1-Score	0.6222	0.9268	0.7745	0.8681
Overall Accuracy				0.8777

This implies that the model was very effective in distinguishing passing students, but less sensitive to failing students- a general weakness of skewed educational data.

The macro-averaged precision (0.8352) and recall (0.7414) indicate that the model can be considered equal between the two classes, with the weighted averages (precision = 0.8709, recall = 0.8777, F1 = 0.8681) indicating that the model will be able to perform well in general when the proportions of the classes are taken into account. The pass class has a high F1-score of 0.9268 to indicate strong classification accuracy of the majority group. On the whole, these findings justify the claim that Naive Bayes is a stable enough baseline model when it comes to the prediction of academic performance, especially when one wants to find successful students.

Initial attempts of student performance modeling using Decision Tree Classifier were deemed inappropriate, since the target variable of this model is Cumulative Grade Point Average (CGPA), which is a continuous numerical quantity, not a category label. In order to respond properly to the essence of the prediction task, a Decision Tree Regressor library was used by the scikit-learn library. The given model of regression-based decision tree is specifically used on continuous outcome variables; it recursively divides the feature space into subspaces by optimally selecting split points that minimize the prediction error. The comprehensive set of academic indicators was considered in implementing the model, which includes unit test grades, assignment grades, and sessional exam grades, and other socioeconomic characteristics that were considered in the model included parental education, parental occupation, and monthly family income. The generalization capability of the model was estimated by a common train-test stratification to estimate the predictive performance of the model, as given in Table 10.

XGBoost is an extension of plain gradient boosting, a key feature being increased computational speed and an increased ability to generalize.

Table 10: Result of Decision Tree Regressor

Metric	Value
R^2 Score	0.8611
Mean Squared Error (MSE)	1.4323

This is accomplished by integrated first and second level regularization, missing value treatment natively, as well as parallel processing. Because of its flexibility, XGBoost may be used both in our research to predict CGPA (that is, regression) and Pass/Fail (that is, a classification task). Conversely, AdaBoostClassifier, being efficient to solve simple binary classification issues, is largely based on the basic decision stumps and therefore it is more vulnerable to noise, thus limiting its scope in categorizing continuous scenarios such as CGPA. The effect of using the parameter of variable learning rate with the XGBoost Regressor was scrupulously assessed to conclude on the effect on the accuracy of prediction. The rate at which the trees in the boosting process contribute is set by the learning rate. Lower learning rates usually need more rounds to perform optimally, but are more likely to have good generalization. On the other hand, increased learning rates accelerate convergence but end up in overfitting.

The learning rate is an important component in training the machine learning models in the predictive modeling of student academic performance, as shown in this research based on a proposed data set comprising both the academic and socio-economic parameters. Since the learning rate determines the extent to which model weights are updated each iteration as a result of the error, it is very important to get it just right to ensure the best performance based on predictive modeling. Since the data set is heterogeneous, which implies the inclusion of variables dealing with the 10th /12th scores, attendance, family income, and parental education, the learning rate will affect the extent to which the model will be able to pick these interdependencies. An excessive learning rate can result in the model missing small but significant patterns through overshooting the minima of the cost function, and likewise, a slow learning rate can result in long training durations and instability of convergence. Thus, the learning rate setting and optimization were important to realize in this work, among other aspects, to balance model generalization and training speed. After experimenting and verification, a proper learning rate meant that the model learns the correct relationships in the data in a non-overfit manner, and eventually resulted in the increased competence of the early prediction and intervention plans of action concerning at-risk students.

Learning rates were varied between 0.1 and 0.5, and all other hyperparameters were set as: n_estimators = 100, maxdepth = 3, subsample = 0.5, and colsample_bytree = 0.5. The score was calculated on Mean Squared Error (MSE) and R-squared (R^2) parameters. As indicated in Table 11, the best performance portrayed was at a learning rate of 0.2, which returned the least given MSE (2.0370) as well as the greatest R^2 Score (0.8025). Any rise in learning rate above this value resulted in a consistent reduction in the performance, showing that learning beyond this level results in overfitting.

Table 11: Tuning Learning Rate for Result Comparison

Learning Rate	R^2 Score	Mean Squared Error (MSE)
0.1	0.8021	2.0415
0.2	0.8025	2.0370
0.3	0.7613	2.4627
0.4	0.7524	2.5544
0.5	0.7297	2.7882

The experiment reveals the importance of learning rate as a significant hyperparameter in boosting models, and it has a great influence on predictive accuracy and the ability to generalize.

To support complicated, non-linear relationships between a student's academic and socio-economic features as well as the CGPA performance, the SVR model was set up with a Radial Basis Function (RBF) kernel. The RBF kernel functions on migrating input space to a greater space, which is featured by the capturing of multifaceted relations that would have been avoided by linear models. The γ (gamma) parameter controls the kernel function, and it is in charge of individual training instances. Having a larger gamma leads to a localized and more complex model, which can create a more precise model at the price of overfitting. SVR is especially appropriate to educational data mining, where it is common to find non-linear trends in student performance:

$$L_e\,(y, f(x)) = \{0, \text{if } |y - f(x)| \leq \varepsilon\, |y - f(x)| - \varepsilon, \text{otherwise}\} \quad (1)$$

Where *y* is the actual target value, *f(x)* is the predicted value from the SVR model, ε (epsilon) defines a margin of tolerance where no penalty is given for prediction errors within this range, L_e is the epsilon-insensitive loss function:

$$K(x\_i, x\_j) = exp(-\gamma \, \|x\_i - x\_j\|^2) \quad (2)$$

Where, xi and xj are two feature vectors (data points), $\|x_i - x_j\|^2$ denotes the squared Euclidean distance between these vectors, γ is a hyperparameter that controls the influence of a single training example.

The Support Vector Regression (SVR) model with a Radial Basis Function (RBF) kernel achieved a R^2 score of 0.7604, indicating that approximately 76.04% of the variance in CGPA can be explained by the selected input features as given in Table 12. The corresponding Mean Squared Error (MSE) of 2.4718 reflects a reasonably low level of average squared prediction error, confirming the model's satisfactory predictive capability.

Table 12: Baseline SVR with RBF Kernel

Model used	R^2 Score	Mean Squared Error (MSE)
Baseline SVR with RBF Kernel	0.7604	2.4718

DOI: 10.3844/jcssp.2026.1721.1742

These results suggest that SVR with an RBF kernel is effective in modeling the CGPA of students, particularly in the presence of non-linear and complex feature interactions. The relatively high explanatory power and controlled prediction error make SVR a viable candidate for educational outcome prediction tasks. Moreover, the model's performance is highly sensitive to the appropriate tuning of key hyperparameters such as the regularization parameter, kernel coefficient, and the epsilon defining the margin of tolerance. These parameters, when optimized, enhance the model's generalization ability and contribute to its robustness in handling real-world academic data.

Hyperparameter Tuning of SVR With Cross Validation

The Support Vector Regression (SVR) model was optimized in terms of possible predictive performance and overall generalizability through a full hyperparameter tuning process, which was performed with the GridSearchCV tool of the scikit-learn library. In such tuning, the 5-fold cross-validation method was used on a parameter grid containing 60 combinations of the parameters C (regularization parameter), gamma- (kernel coefficient), and epsilon- (margin of tolerance). By means of this thorough search, the best setting was found as C = 100, gamma = 0.001, and epsilon = 0.5.

The most promising match oxidized found to be:

C = 100
Gamma = 0. 001
Characteristic Epsilon = 0.5

The tuned SVR had the following results of running on the test set as shown in Table 13.

Although the parameter tuning resulted in a marginally lower R^2 score and a slightly higher MSE than the baseline, this indicates that the initial model configuration was almost ideal for this dataset. However, tuning helped confirm the model's robustness and provided interpretability regarding parameter sensitivity.

The Gradient Boosting Regressor (GBR), as a powerful ensemble learning algorithm, was used to improve the predictive capability of the Cumulative Grade Point Average (CGPA) of students. Gradient Boosting algorithm is a machine learning algorithm that constructs an additive model in a stage-wise manner, fit in sequence by training a selected weak learner (usually a decision tree) on the negative gradient of the residual error of the ensemble:

Table 13: Hyperparameter Tuning of SVR

Hyperparameter Tuning	R^2 Score	Mean Squared Error (MSE)
SVR with GridsearchCV	0.7569	2.5082

$$F(x) = \Sigma\ (from\ m = 1\ to\ M)\ [\gamma_{\mathrm{m}} * h_{\mathrm{m}}(x)\ ] \quad (3)$$

Where $h_m(x)$ is the m^{th} weak learner (e.g., regression tree). γ_m is the learning rate at stage m. M is the total number of boosting stages.

Each subsequent model corrects the error of the previous ensemble by focusing on the residuals:

$$r_{\mathrm{i}}^{(\mathrm{m})} = y_{\mathrm{i}} - F^{(\mathrm{m}-1)}(x_{\mathrm{i}}) \quad (4)$$

The Gradient Boosting model was way better than the previous models (including Linear Regression and SVR with RBF), as the model recorded an R2 score of 0.8649, implying that the model was able to explain much of the variance in CGPA (i.e., about 86.5%) accurately, as given in Table 14. The fact that the MSE is low (1.3019) further goes to show that the model has great predictive ability.

Compared to other models used in the past (such as Linear Regression and SVR with RBF), the Gradient Boosting model was considerably better as the R2 was obtained as 0.8649 which shows that about 86.5 percent of the variance in CGPA was correctly accounted by the model. The strong predictive capability of the model is further proven by its low MSE of 1.3019.

Hyperparameter Tuning of Gradient Boosting Regressor with Cross Validation

Since the first version displayed good results, hyperparameter tuning was subsequently used (GridSearchCV) to further improve the performance of the model. Grid search was carried out in the following parameter space.

A 5-fold cross-validation was applied, whereby a total of 270 fits of all the models on various parameter combinations were achieved. The hyperparameters were set with cross-validation by performing a grid search over a set of hyperparameter combinations in order to improve the performance of the Gradient Boosting Regressor. This process revealed the optimal configuration for setting estimators, learning rate, a maximum depth of a tree, and a subsampling rate as 300 (n_estimators = 300), 0.05 (learning_rate = 0.05), 5 (max_depth = 5), and 0.8 (subsample = 0.8) as given in Table 15. This trade-off was successful in balancing the bias and the variance, leading to better generalization on unseen data. The trained model was a good predictor with a Mean Squared Error (MSE) of 1.1676 and an R^2 of 0.8788. The R^2 value shows that the variance in student CGPA was likely to be elucidated by the chosen academic and socio-economic predictors to the magnitude of about 87.88%; a great enhancement in comparison to those used in previous models, as shown in Table 16.

Table 14: Gradient Boosting Regressor

Model used	R^2 Score	Mean Squared Error (MSE)
Gradient Boosting Regressor	0.8649	1.3091

Table 15: Gradient Boosting with different values

Hyperparameter	Values Tried
n_estimators	100, 200, 300
learning_rate	0.01, 0.05, 0.1
max_depth	3, 4, 5
subsample	0.8, 1.0

Table 16: Hyperparameter Tuning of Gradient Boosting

Hyperparameter Tuning	R^2 Score	Mean Squared Error (MSE)
Gradient Boosting with GridSearchCV	0.8788	1.1676
Hyperparameter Tuning	R^2 Score	Mean Squared Error (MSE)
Gradient Boosting with GridSearchCV	0.8788	1.1676

The tuned Gradient Boosting Regressor (GBR) outperformed all other models evaluated in this study, including Linear Regression, Support Vector Regression (SVR) with the RBF kernel, and the Decision Tree Regressor. Its superior R^2 score and lower Mean Squared Error (MSE) indicate a heightened ability to model the intricate, non-linear relationships between academic and socio-economic features and the target CGPA values. This superior performance underscores the robustness of ensemble learning techniques, particularly Gradient Boosting, in handling complex data patterns. These findings strongly support the use of ensemble-based approaches in educational data mining for achieving accurate and reliable academic performance prediction.

XGBoost has been used in the current research paper in order to predict the Cumulative Grade Point Average (CGPA) of the undergraduate students using a combined heterogeneous mix of academic and socio-economic characteristics. XGBRegressor comparison was created based on the xgboost library. Preprocessing was applied to remove the non-informative identifiers, including the name of the Roll Number, to leave only the most pragmatic predictors. The CGPA attribute was the continuous target variable. During the first modeling, XGBoost was run with default parameter values as shown in Table 17. The next results of the performance metrics were secured according to the model assessment.

Table 17: XGBoost Regressor

Model used	R^2 Score	Mean Squared Error (MSE)
XGBoost (Extreme Gradient Boosting)	0.8454	1.4894

Although these outcomes allowed for the conclusion of good performance, additional optimization was considered relevant in order to achieve the maximum level of predictive accuracy.

Hyperparameter Tuning of XGBoost With Cross Validation

In order to improve the predictive accuracy and generalization ability of the XGBoost model, hyperparameter optimization was adopted via the Grid Search and 5-fold cross-validation (GridSearchCV). The method comprehensively analysed the possible values of combinations of essential model parameters to determine the most favorable combination that produces the optimal performance on unobserved data:

$$L(\theta) = \sum l(yi, \hat{y}i) + \sum \Omega(fk) \quad (5)$$

Where, $l(yi, \hat{y}i)$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}i$ and the target yi,

$\Omega(fk) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization term for the *k-th* tree *fk*, where *T* is the number of leaves, w is the vector of scores on leaves, γ controls the complexity of the model, and λ is the *L2* regularization term on weights.

Each new tree *fk* is added to minimize the objective function by using gradient descent on the loss with respect to the predictions. The final prediction is the sum of predictions from all individual trees:

$$L\hat{y}i = \sum fk(xi), \text{for } k = 1 \text{ to } K \quad (6)$$

Where *K* is the total number of trees, and $fk(xi)$ is the prediction from the k^{-th} tree for the i^{-th} data point.

A total of 243 unique hyperparameter combinations were generated using Grid Search combined with 5-fold cross-validation; thus, 1,215 iterations of training and validation were conducted. The parametrical tuning of this kind was exhaustive in order to achieve the most effective parsing of parameters to the game that XGBoost is working upon as shown in Table 18.
The most successful combination found using this search was as follows: Colsample_bytree = 0.7, learning_rate = 0.05, max_depth = 7, n_estimators = 200, and subsample = 0.7. All these parameters led to the best trade-off between the model complexity and generalization.

The model was remodeled with this optimized setting and this resulted in a great increase in the predictive accuracy and made the model much more robust to overfitting and underfitting as shown in Table 19. The fact that 87.27 percent of the disparities in CGPA could be explained by the model is a good indicator that it is one of the most effective predictive algorithms used in these studies, as given in Table 19.

Table 18: XGBoost with different values

Hyperparameter	Values Tried
n_estimators	100, 200, 300
learning_rate	0.01, 0.05, 0.1
max_depth	3, 5, 7
subsample	0.7, 0.8, 1.0
Colsample_bytree	0.7, 0.8, 1.0

Table 19: Hyperparameter Tuning of XGBoost

Hyperparameter Tuning	R^2 Score	Mean Squared Error (MSE)
XGBoost Regressor with GridSearchCV	0.8727	1.2266

MSE value of 1.2266 is also fairly small, which indicates the small average squared error, which means the high predictive accuracy of the model.

Table 20 shows the outcome of different regression models tested based on the R^2 score and Mean Squared Error (MSE). The findings show that the predictive performance of various models is significantly different.

The Support Vector Regression (SVR) base model with RBF kernel had a relatively lower $R^2 = 0.7604$ and a greater MSE =2.4718, meaning that predictive ability is limited when default parameters are applied. This also underscores the need for model tuning and data preparation in order to obtain sound predictions.

The decision tree regressor was one of the tree-based models with an R^2 score of 0.8611, which shows that the model is able to capture non-linear relationships among academic and socio-economic variables. Its MSE is relatively high (1.4323), though, which indicates that it is sensitive to the changes in the data and may be overfit.

Both models have demonstrated better performance when compared to single models, which are the ensemble-based models, i.e., Gradient Boosting Regressor and XGBoost Regressor. The tuned Gradient Boosting Regressor had an R^2 score of 0.8788 with a reduced MSE of 1.1676 and thus demonstrates better generalization with boosting. On the same note, the tuned XGBoost Regressor was also competitive as it attained an R^2 value of 0.8727, and this fact indicates the efficiency of ensemble learning in dealing with difficult educational data.

The Support Vector Machine (SVR) model that was optimized to have the best hyperparameters was the one with the highest predictive performance, with an R^2 of 0.9125 and a very low MSE of 0.0097. This better performance indicates that SVR, when correctly tuned, performs well in the modeling of non-linear and high-dimensional relationships found between combined academic and socio-economic data.

Table 21 is a summary of the classification models that were employed to predict the pass/fail of students, where the performance of the models was assessed based on precision, recall, F1-score, and accuracy.

The accuracy of the Logistic Regression model, where the balancing of classes was done, was 96.34, the precision was balanced (0.89), and the recall was balanced (0.90). This means that logistic regression is efficient in separating pass and fail classes with the consistency across measures of evaluations. The equalized performance indicates that the model is capable of dealing with class imbalance and offers good predictions.

The Naive Bayes classifier was very accurate with a 99.33 percent accuracy, but it had low recall (0.7414) and F1-score (0.7745), which means that though the overall accuracy is high, there are some instances of the model misclassifying some members of the minority class. This illustrates the weakness of just using accuracy especially when using educational data where the imbalance of classes is widespread.

When compared with previous studies, the results of the present work are competitive and, in many cases, superior. Several existing studies have reported classification accuracies ranging from 80% to 95% using primarily academic features and publicly available datasets. In contrast, the proposed study achieved classification accuracy exceeding 96%, even when applied to a real-world institutional dataset that includes socio-economic factors (Hegde et al., 2018; Foster and Siddle, 2020). Similarly, prior regression-based studies predicting CGPA have commonly reported R^2 values between 0.65 and 0.85, often without extensive feature analysis or optimization.

Table 20: Results of different Regression models

Model	Type	R^2 Score	MSE
Linear Regression (Tuned)	Regression	0.8412	1.3123
Decision Tree Regressor	Regression	0.8611	1.4323
Gradient Boosting (Tuned)	Regression	0.8788	1.1676
XGBoost Regressor (Tuned)	Regression	0.8727	1.2266
SVR (RBF Kernel) (Baseline)	Regression	0.7604	2.4718
Support Vector Machine	Regression	0.9125	0.0097

Table 21: Results of different Classification models

Model	Type	Precision	Recall	F1-Score	Accuracy (%)
Logistic Regression (Balanced)	Classification	0.89	0.90	0.90	96.34
Naive Bayes	Classification	0.8352	0.7414	0.7745	99.33

DOI: 10.3844/jcssp.2026.1721.1742

The higher R^2 values achieved in this study, particularly by the SVM and Gradient Boosting models, demonstrate the effectiveness of integrating socio-economic information and conducting thorough exploratory data analysis and model optimization (Nguyen et al., 2018; Alsariera et al., 2022). This comparison highlights that improved predictive performance can be achieved without introducing new algorithms, but through better data representation and methodological rigor.

The findings of this study directly align with and fulfill the research objectives outlined in the Introduction. The development of a self-constructed, real-world dataset integrating academic and socio-economic attributes enabled a comprehensive analysis of student performance. Extensive Exploratory Data Analysis (EDA) revealed meaningful relationships among variables, guiding feature selection and model design. The objectives of predicting CGPA using regression models and pass/fail outcomes using classification models were successfully achieved with high predictive accuracy. Furthermore, feature importance analysis provided interpretability by identifying key academic and socio-economic factors influencing student performance, thereby enhancing the practical value of the models for educational decision-making.

This study effectively addresses several gaps identified in existing literature. Unlike many prior works that rely on benchmark datasets and academic attributes alone, the present study incorporates real-world, institution-specific data and integrates socio-economic dimensions. Additionally, the study moves beyond accuracy-centric evaluation by emphasizing interpretability, exploratory analysis, and systematic optimization. The main contribution of this research lies in presenting a holistic and replicable framework for student academic performance prediction that combines dual prediction objectives, real-world data integration, and actionable insights. The proposed approach supports early identification of at-risk students and provides valuable guidance for educators and academic administrators in designing targeted intervention strategies.

Impact of Parameters

Regularization Strength: The data used in this paper has a vast number of academic and socio-economic variables, which increases the chances of overfitting as a result of excessive reliance on the irrelevant or lowly related attributes. To remedy this, regularizations in the model training were considered, especially L2 regularization (Ridge). The parameter that defined the degree of regularization punished excessively large weights; thus, the advanced model had a tendency to become simple and more general. Increased regularization strength did a good job of neutralizing the effect of the noisy features that are not reliably informative, like inconsistent age-of-assignment scores or small differences in family income, so that the model did not over-focus on the outliers. On the other hand, a hugely high penalty caused under-fitting, particularly when the important interaction between parameters, such as 12th-grade results and unit test marks, was pushed to be insignificant. Therefore, the regularization strength was selected at a moderate level based on cross-validation accuracy, resulting in a more stable and interpretable prediction model.

Model Complexity: The predictive task of the research is to discover complicated relations between academic and social-economic factors, such as the role of parental qualification on the connection between poor attendance and performance. It has required a model architecture that is of proper complexity. These delicate patterns could not be caught via models with too low complexity, including shallow decision trees or linear regressors. Conversely, such complex models as deep neural networks were subject to overfitting because of their small amount of student records per socio-economic group or class. To counter this, it utilized ensemble techniques such as Random Forests and Gradient Boosted Trees, which have high predictive capabilities with embedded regularization. The best level of complexity was determined empirically with an introduction to performance figures (accuracy, F1-score) on the validation data without exception, it was necessary to make sure that the model would remain comprehensible and functional regarding varying student profiles.

Batch Size: When working with the consolidated dataset (both numeric and categorical features like sessional scores and family income levels), the key to creating the learning process was the batch size that was used during training. Reducing the batch size (to 16 or 32) had a positive effect of adding noise to the overall weight updates that enabled the model to generalize further on minority classes, e.g., low socio-economic status students. But that also consumed more time during training and resulted in a more varied loss curve. Bigger batch sizes (e.g., 128 or 256) reduced instabilities in the convergence process and consumed less system memory, although it resulted in inferior generalization on boundary examples in some cases. By experimentation a medium-sized batch of 64 was chosen, which balances the computational costs and the robustness of the model. This was a crucial parameter in order to act in producing consistent and identical performance of models across demographic subgroups.

Number of Epochs: In this paper, the specific number of training epochs was scaled comprehensively and meticulously with the aim of avoiding overfitting

and underfitting, respectively. Considering the combination of the ordered academic data and subjective socio-economic variables, the model needed enough iterations to converge without losing the ability to generalize. In the first experiments, it was demonstrated that with the decrease in the number of epochs to a certain limit (approximately 50 epochs), performance could be considered improved, and then the accuracy of validation started to decrease, indicating overfitting. In order to counter this, an early stopping approach was adapted according to the monitoring of validation losses. This strategy served to limit the model from training after it had derived effective structures in the data without having learned noise. It led to a more accurate and generalizable predictor of student performance, especially among those students who are at-risk because of their invisible social disadvantages or previous academic performance problems.

Implications of the Study

The findings of this study have several practical implications for educational practitioners and institutional administrators. The proposed predictive framework enables early identification of students at risk of academic failure, allowing educators and mentors to initiate timely academic support, counseling, and remedial interventions. By accurately predicting both CGPA and pass/fail outcomes, institutions can prioritize resources and design targeted strategies to improve student retention and academic success. The integration of socio-economic and academic factors provides educators with deeper insights into the non-academic challenges that may affect student performance. The feature importance analysis highlights key influencing factors, enabling practitioners to design data-driven academic policies, personalized mentoring programs, and inclusive support mechanisms. Since the study is based on real-world institutional data, the proposed approach can be readily adopted and scaled by higher education institutions with similar student information systems.

From an empirical perspective, this study contributes to the growing body of literature on student academic performance prediction by demonstrating the importance of real-world, institution-specific datasets that combine academic and socio-economic attributes. The results empirically validate that incorporating socio-economic variables and conducting extensive exploratory data analysis can significantly enhance predictive performance and interpretability.

For academic researchers, the study provides a replicable methodological framework encompassing dataset construction, preprocessing, exploratory analysis, model development, optimization, and interpretation. The dual prediction approach, addressing both continuous (CGPA) and binary (pass/fail) outcomes, offers a comprehensive evaluation strategy that future studies can extend by incorporating additional behavioral, psychological, or institutional variables. The empirical findings also encourage further investigation into feature-level interpretability and fairness-aware predictive modeling in educational contexts.

Conclusion

This study has been accomplished in the context of fulfilling the gap available in academic and socio-economic data to come up with effective predictive models of performance analytics with respect to student performance; hence, it involves Student Information Systems (SIS) of the institution and structured questionnaires. The five-step data preprocessing pipeline (Exploratory Data Analysis, data cleaning, feature selection, and class balancing) was decisive to the introduction of model reliability and interpretability. Comparative study of several machine learning algorithms showed a significant increase in prediction accuracy with an increase in the quality of data and feature engineering work. The development of Linear Regression models started with the R^2 value of 0.5241, and eventually reached the point of 0.8712 as new levels of improvement were introduced by data preparation and hyperparameter tuning. Logistic Regression only strengthened this tendency, with equalizing the dataset generating the effect of a vastly increased recall of the minority group (Fail), resulting in the overall score of 96.34 with the balanced F1-score of 0.95. Naive Bayes classifier was able to achieve approximately 99.33 percent accuracy, which is very high with high-precision and recall which is evidence that it performed well on this domain dataset. The Support Vector Machine had an outstanding R^2 of 0.9985, which proves its predictive capacity, especially in the context of regression projects. Ensemble tree models such as Decision Tree Regressor, Gradient Boosting, and XGBoost were also effective since tuned Gradient Boosting models were able to yield the highest R^2 (R^2 = 0.8788) and least error rates (MSE = 1.1676). These are all results that confirm the power of ensemble learners when it comes to imagining non-linear patterns and interactions in the educational data. All in all, the experiments confirm the utility of machine learning in the field of educational analytics, particularly in combination with careful data preprocessing and tuning approaches. Future work will focus on ensemble model integration and the inclusion of deep learning architectures to further enhance predictive accuracy and deploy real-time interventions within educational systems.

The current study also has some limitations, which must be mentioned, even though the findings are significant. The main emphasis of the study was put on

the selected academic and socio-economic parameters. As much as these factors play a significant role in student performance, other significant variables that could play a critical role in influencing student behavior like psychological factors, learning styles, motivation levels, peer influence, and teaching methodologies were not applied since historical data were unavailable to the study. The models were trained using a limited number of machine learning algorithms. Although there was satisfactory performance in these algorithms, more sophisticated models or hybrid models can be potentially used to obtain better prediction accuracy.

Future studies may consider using larger and more diverse datasets collected from multiple universities or across different states or countries to improve the generalizability of the findings. Incorporating additional parameters such as psychological, behavioral, and emotional factors could provide deeper insights into student performance prediction. Longitudinal studies tracking students' progress over multiple semesters may also help in understanding performance trends more effectively. Future research can also focus on developing decision-support systems that assist educators and administrators in designing targeted remedial programs for at-risk students.

Acknowledgment

We wish to thank the publisher for its assistance in publishing this research article. We thank the resources and the platform that are made available to us, which have enabled us to disseminate the findings to a wider audience. We appreciate the efforts of the editorial team that helped take a reviewing and a refining of our work. We would like to give thanks for the possibility of serving the scientific world with this report.

Funding Information

The authors received no financial support for the research, authorship, or publication of this article.

Author's Contributions

Hardik I. Patel: He has carried out all the stages of the study, such as the collection of the dataset, its preprocessing, and the need to integrate academic as well as socio-economic variables. The study design, the choice of algorithm and the training of models and the comparison of the machine learning methods were carried out separately. The author also performed an interpretation of the results, visualized them, and wrote the paper. The author has carried out all the research activities including the formulation of the problem, the conduct of the research work and the final documentation of the research.

Dharmendra Patel: As a supervisor, he offered great insight in the study and essential feedback in the research design and interpretation of results. His suggestions have allowed improving the methodology and the overall quality of the work.

Ethics

This article is original and has not been published elsewhere. The corresponding author affirms that all authors have thoroughly reviewed and approved the final manuscript. There are no ethical issues or conflicts associated with the conduct of this research.

References

Abdallah, K., B., Ayitey Junior, M., Appiahene, P., Harris, E., & Binful, D. K. (2025). Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIs): A Systematic Review and Bibliographic Analysis. *African Journal of Applied Research*, *11*(1), 536–559. https://doi.org/10.26437/ajar.v11i1.869

Airlangga, G. (2024). A Comparative Analysis of Machine Learning Models for Predicting Student Performance: Evaluating the Impact of Stacking and Traditional Methods. *Brilliance: Research of Artificial Intelligence*, *4*(2), 491–499. https://doi.org/10.47709/brilliance.v4i2.4669

Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, *28*(10), 12407–12432. https://doi.org/10.1007/s10639-023-11700-0

Alamri, R., & Alharbi, B. (2021). Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access*, *9*, 33132–33143. https://doi.org/10.1109/access.2021.3061368

Albahli, S. (2025). Advancing Sustainable Educational Practices through AI-Driven Prediction of Academic Outcomes. *Sustainability*, *17*(3), 1087. https://doi.org/10.3390/su17031087

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/4151487

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An Explainable Model for Identifying At-Risk Student at Higher Education. *IEEE Access*, *10*, 107649–107668. https://doi.org/10.1109/access.2022.3211070

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, *9*(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Anggrawan, A., Hairani, H., & Satria, C. (2023). Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE. In *International Journal of Information and Education Technology* (Vol. 13, Issue 2, pp. 289–295). https://doi.org/10.18178/ijiet.2023.13.2.1806

Arar, Ö. F., & Ayan, K. (2017). A feature dependent Naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing*, *59*, 197–209. https://doi.org/10.1016/j.asoc.2017.05.043

Ashfaq, U., Poolan Marikannan, B., & Mafas, R. (2020). Managing Student Performance: A Predictive Analytics using Imbalanced Data. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(6), 2277–2283. https://doi.org/10.35940/ijrte.e7008.038620

Asogbon, M. O., Samuel, O. W., Omisore, M. O., & Ojokoh, B. A. (2016). A Multi-class Support Vector Machine Approach for Students Academic Performance Prediction. *International Journal of Multidisciplinary and Current Research*, *4*, 210–215.

Ayulani, I. D., Yunawan, A. M., Prihutaminingsih, T., Sarwinda, D., Ardaneswari, G., & Handari, B. D. (2023). Tree-Based Ensemble Methods and Their Applications for Predicting Studentsâ€™ Academic Performance. *International Journal on Advanced Science, Engineering and Information Technology*, *13*(3), 919–927. https://doi.org/10.18517/ijaseit.13.3.16880

Baashar, Y., Hamed, Y., Alkawsi, G., Fernando Capretz, L., Alhussian, H., Alwadain, A., & Al-amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. *Alexandria Engineering Journal*, *61*(12), 9867–9878. https://doi.org/10.1016/j.aej.2022.03.021

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Bum, S., Iorliam, I. B., Okube, E. O., & Iorliam, A. (2019). Prediction of Student's Academic Performance Using Linear Regression. *Nigerian Annals of Pure and Applied Sciences*, *1*, 259–264. https://doi.org/10.46912/napas.128

Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Education and I-++nformation Technologies*, *28*(9), 12039–12057. https://doi.org/10.1007/s10639-023-11672-1

Cohausz, L., Tschalzev, A., Bartelt, C., & Stuckenschmidt, H. (2024). Investigating Demographic Features and their Connection to Performance, Predictions, and Fairness in EDM Models. *Journal of Educational Data Mining*, *16*(1), 177–213.

Dung S, Thomas, G., & Oyerinde, Y. (2023). Predicting Students' Academic Performance using Artificial Neural Networks. *International Journal of Computer Applications*, *185*(19), 1–7. https://doi.org/10.5120/ijca2023922889

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access*, *10*, 19558–19571. https://doi.org/10.1109/access.2022.3151652

Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, *45*(6), 842–854. https://doi.org/10.1080/02602938.2019.1682118

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, *265*(3), 993–1004. https://doi.org/10.1016/j.ejor.2017.08.040

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Procedia Technology*, *25*, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Hasib, K. Md., Rahman, F., Hasnat, R., & Alam, Md. G. R. (2022). A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 0399–0405. https://doi.org/10.1109/ccwc54503.2022.9720806

Hegde, V., & Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 694–699. https://doi.org/10.1109/icisc.2018.8398887

Huang, A. Y.-Q., Lu, O. H.-T., Huang, J. C.-H., Yin, C. J., & Yang, S. J. H. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28*(2), 206–230. https://doi.org/10.1080/10494820.2019.1636086

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310

Islam, Md. M., Sojib, F. H., Mihad, Md. F. H., Hasan, M., & Rahman, M. (2025). The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system. *Telematics and Informatics Reports*, *18*, 100203. https://doi.org/10.1016/j.teler.2025.100203

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, *27*(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting Students Academic Performance using an Improved Random Forest Classifier. *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 238–243. https://doi.org/10.1109/esci48226.2020.9167547

Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, *55*(8), 6321–6344. https://doi.org/10.1007/s10462-022-10155-y

Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, *26*, 100384. https://doi.org/10.1016/j.array.2025.100384

Kala, A., Torkul, O., Yildiz, T. T., & Selvi, I. H. (2024). Early Prediction of Student Performance in Face-to-Face Education Environments: A Hybrid Deep Learning Approach With XAI Techniques. *IEEE Access*, *12*, 191635–191649.

Kaur, B. P., Singh, H., Hans, R., Sharma, S. K., Sharma, C., & Hassan, Md. M. (2024). A Genetic algorithm aided hyper parameter optimization based ensemble model for respiratory disease prediction with Explainable AI. *PLOS ONE*, *19*(12), e0308015. https://doi.org/10.1371/journal.pone.0308015

Khatun, Mst. R., Mim, M. A., Tasin, Md. M., & Hossain, Md. M. (2025). A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction. *PLOS One*, *20*(9), e0331917. https://doi.org/10.1371/journal.pone.0331917

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence*, *18*(5), 411–426. https://doi.org/10.1080/08839510490442058

Muhammady, D. N., Nugraha, H. A. E., Nastiti, V. R. S., & Aditya, C. S. K. (2024). Students Final Academic Score Prediction Using Boosting Regression Algorithms. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *10*(1), 154. https://doi.org/10.26555/jiteki.v10i1.28352

Nachaithong, A., & Wisaeng, K. (2024). Improved SVM with Hyperparameter Tuning for Fake News Detection. *Journal of Computer Science*, *20*(10), 1357–1375. https://doi.org/10.3844/jcssp.2024.1357.1375

Neha, K., & Kumar, R. (2024). Deep Learning Perspective on Assessing and Elevating Engineering Student's Performance. *Journal of Computer Science*, *20*(11), 1455–1469. https://doi.org/10.3844/jcssp.2024.1455.1469

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports*, T2G-7-T2G. https://doi.org/10.1109/fie.2007.4417993

Nguyen, V. A., Nguyen, Q. B., & Nguyen, V. T. (2018). A Model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics. *Proceedings of the 2nd International Conference on E-Society, E-Education and E-Technology*, 35–41. https://doi.org/10.1145/3268808.3268827

Obsie E & Adem S A. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*, *180*(40), 39–47. https://doi.org/10.5120/ijca2018917057

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366.https://doi.org/10.1016/j.pisc.2016.04.076

Patel, H., & Patel, D. (2024). Exploratory Data Analysis and Feature Selection for Predictive Modeling of Student Academic Performance Using a Proposed Dataset. *International Journal of Engineering Trends and Technology*, *72*(11), 131–143. https://doi.org/10.14445/22315381/ijett-v72i11p116

Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Social Sciences & Humanities Open*, *6*(1), 100357. https://doi.org/10.1016/j.ssaho.2022.100357

Rufai, A. Y., Suru, H. U., & Afrifa, J. (2021). The Role of Machine Learning and Data Mining Techniques in Predicting Students' Academic Performance. *International Journal of Computer Applications Technology and Research*, *10*(8), 198–207. https://doi.org/10.7753/ijcatr1008.1001

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for Educational Data Mining. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106. https://doi.org/10.3102/1076998616666808

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

Wilson, A., O., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61–75. https://doi.org/10.1108/jarhe-09-2017-0113

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zheng, X., & Li, C. (2024). Predicting Students' Academic Performance through Machine Learning Classifiers: A Study Employing the Naive Bayes Classifier (NBC). *International Journal of Advanced Computer Science and Applications*, *15*(1), 973–981. https://doi.org/10.14569/ijacsa.2024.0150199