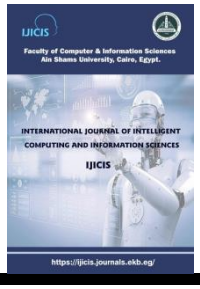
IJICIS, Vol.25, No.3, 55-72
DOI: 10.21608/ijicis.2025.411771.1417

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



MACHINE LEARNING-BASED PREDICTIVE MODELING OF STUDENT ACADEMIC PERFORMANCE

Israa Mohamed*
Faculty of Computers and Informatics, Zagazig University,
Zagazig 44519, Egypt
Faculty of Engineering and Computer Sciences, King Salman International University,
South Sinai, Egypt
israa.mohamed2222@gmail.com
Received 2025-08-07; Revised 2025-08-30; Accepted 2025-09-01

Abstract *Towards predicting university student academic success five machine learning algorithms (Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks) will be evaluated for their performance effectiveness in this study. The research investigates which predictive modelling technique provides maximum reliability for data-driven choices within higher education institutions. The research evaluated different machine learning methods when they analyzed educational datasets. The evaluation of models utilized accuracy precision along with specificity and the F1 score for metric assessment. A systematic testing method was used during model training and testing to establish reliable results for every algorithm. Neural Networks produced the most effective results with 69.83% accuracy and 80.29% F1 score, yet Random Forest achieved 69.27% accuracy combined with 80.09% F1 score. The accuracy measures of Support Vector Machines and Logistic Regression reached 69.27%, but Decision Trees produced 65.88% accuracy. Educational data analysis benefited the most from complex models when these models surpassed simpler algorithms in identifying complex associations. The research establishes an inclusive evaluation of different machine learning applications on educational data which addresses a literature gap about predicting methods in academic environments. Based evidence supports institutions wishing to implement predictive analytics systems for student performance monitoring.*

Keywords: Academic performance prediction, Machine learning, Educational data mining, Predictive analytics

1. Introduction

The education sector uses big data analytics as an unmatched chance to enhance academic institutional results and student achievement metrics [1]. The digital revolution directly affects higher education because student data volumes and varieties are expanding rapidly [2], [3]. Educational institutions in the

*__Corresponding Author__: Israa Mohamed

Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt

Faculty of Engineering and Computer Sciences, King Salman International University, South Sinai, Egypt, Cairo, Egypt

Email address: israa.mohamed2222@gmail.com

United States encounter growing expectations to raise their student retention percentages and academic outcomes and success results [4]. Academic performance predictive modeling takes interest due to increased pressure and advanced analytical capabilities [5]. Educational organizations are moving toward surpassing their reliance on conventional descriptive data collection by implementing more advanced predictive methodologies [6], [7].

The key requirement involves building dependable prediction systems that properly connect the multifaceted elements that impact students' performance, including their previous academic results, demographic profile, socioeconomic standing, and behavioral tendencies. Conventional statistical tools have shown value by providing insights but cannot fully portray complex variable interactions [8].

The research tackles this problem by evaluating different machine learning algorithms that predict university student academic performance levels. Five major models, including Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Networks, are analyzed as part of our evaluation. This study seeks to identify the best predictive modeling solution for student performance assessment by systematically examining different algorithms regarding their multiple performance indicators. The study focuses on model comparison because educational organizations require proof-based recommendations for adopting predictive analytics systems [9].

The research results show different performance outcomes among the analyzed models. A Neural Network achieved the best results by reaching 69.83% accuracy and 80.29% F1 score, but Random Forest stayed close with 69.27% accuracy and 80.09% F1 score. The performance amounts between Support Vector Machines and Logistic Regression were equal, with 69.27% accuracy, yet Decision Trees scored 65.88% accuracy. The data patterns in educational data seem better characterized through Neural Networks and Random Forests rather than simpler techniques.

The findings provide essential guidance to educational institutions using predictive analytics systems [10]. Academic institutions should consider investing in advanced analytical systems because ensemble and deep learning approaches show superior outcomes, but logistic Regression is a dependable baseline solution [11]. The research follows a specific structure in which Section 2 surveys educational data mining and predictive analytics literature, Section 3 explains methodology along with data collection descriptions, Section 4 shows results and analyzes them, and the final part of Section 5 explores limitations and implications with future research recommendations.

2. Literature Review

2.1 Educational Data Mining and Predictive Analytics

Educational Data Mining (EDM) utilizes data mining and machine learning techniques to analyze educational data, aiming to enhance learning outcomes and support statistical analysis. [12]. The core objective of (EDM) involves obtaining significant patterns from big datasets so educators and administrators receive actionable insights. The field focuses on student risk identification at early stages because this enables necessary corrections leading to better academic performance [13], [14].

2.2 Factors Influencing Student Performance

Multiple research findings have identified various related elements that strongly influence how students perform academically. These elements exist throughout different aspects of student environmental factors. Academic factors encompass traditional metrics such as grades, attendance records, and participation in extracurricular activities. Student academic performance depends heavily on age, gender

identity, socioeconomic situation, and their parents' educational attainment [14]. Behavioral components examine the practices of studying, students' online platform interaction, and their quality relationships with instructors and classmates. According to [15], combining academic and non-academic features proves more effective for student performance prediction

2.3 Machine Learning Models for Student Performance Prediction

Educational research implements the baseline model of Logistic Regression because of its straightforward nature and easy interpretation appeal to researchers. The findings by [16] established that simple datasets generate good performance from Logistic Regression, yet these methods are ineffective for advanced educational problems and nonlinear relationships. Logical Regression proves worthwhile for educational data mining because it generates precise probability measures and readable coefficients when measuring predictor variables. Logistic Regression can create outcomes that match complex prediction models' performance when implemented alongside regulated feature selection and preprocessing steps. It is also used to predict educational binary outcomes. Its explainability features make it highly beneficial to organizations that need transparent prediction systems that understand computational models' rationale [17].

The broad adoption of decision trees stems from their commitment to controlling nonlinear data relationships alongside their precise results interpretation capabilities. The systematic partitioning methods used by these models enable them to standalone vital factors that affect student performance. [18] established Decision Trees outperform traditional statistical methods for predicting academic success because they combine flexible features with user-friendly interfaces [19]. In addition, a similar method approach was conducted to compare the difference approach between the traditional statistical approach and machine learning to predict cancer patients [20].

The visual nature of Decision Trees as decision-making tools serves educational practitioners exceptionally well by presenting them with an understandable way to represent these processes [21]. The visualization capability of Decision Trees allows educational administrators, together with teachers, to see which factors significantly impact student performance, starting with attendance numbers through engagement variables [22]. The research demonstrated how Decision Trees use binary splitting rules to identify meaningful patterns that educational institutions can convert into precise support programs [23]. Predictive accuracy increases through Random Forest ensemble learning because it joins several Decision Trees to reduce overfitting risks[24]. Random Forest has demonstrated exceptional performance with complex dataset interactions. The model's ability to rank important features delivers valuable information that helps educators identify vital performance elements [25].

A critical factor that strengthens Random Forest's application in educational research is its disappearance algorithm, better known as bootstrap aggregating (bagging) [14]. The researchers proved that this methodology stands out in managing unbalanced education datasets, which include at-risk student identification with low representative struggling students compared to typical students. Random Forest demonstrates through research that it combines ensemble learning with reliable performance measurements that yield stable estimates across academic environments while improving prediction accuracy [26].

SVM demonstrates exceptional performance when dealing with high-dimensional datasets since they have become essential for student success predictions in both online education and blended learning settings [27]. The research of [28] showed how SVM demonstrated an exceptional ability to forecast

student departure in Massive Open Online Courses (MOOCs) with feature selection methods available for optimization.

According to [29], the sophisticated mathematical foundation of SVM offers unique advantages in educational data analysis through its kernel-based approach. Different kernel functions enable the detection of diverse relationships between educational data points, ranging from straightforward linear patterns in schoolroom measures to advanced nonlinear behaviors in online student activities. SVM more easily recognizes student engagement patterns and performance information due to its space transformation capabilities, which detect patterns that basic algorithms would miss, especially in hybrid educational settings that produce numerous educational data types [30].

Educational data mining now uses XGBoost as a robust gradient-boosting algorithm that has proven effective in predictive modeling tasks. The ensemble method operates beyond traditional boosting with added regularization elements that suppress overfitting. XGBoost delivered extraordinary results for predicting tasks when processing data sets from education with integrations between multiple feature varieties [31]. The scalability combined with missing values handling features of this algorithm ensures its suitability for large educational data analysis projects.

As one of the cutting-edge prediction techniques for student performance, Neural Networks detect intricate nonlinear patterns found within educational data sets. Neural Networks can process varied educational datasets incorporating regular academic metrics and irregular data sets such as student behavioral patterns. Neural Networks show an advantage in performance identification by noticing subtle student behavior patterns that conventional models would overlook [32]. The self-learning abilities of these models empower their valuable use in educational scenarios, which contain complex relationships between variables that human analysts do not detect easily [33].

2.4 Comparative Studies, Research Findings and Gaps

Various machine learning models receive an ongoing comparative evaluation by the field to determine their effectiveness in student performance predictions [34]. A thorough assessment by [35] showed ensemble techniques, especially Random Forest, outperformed Logistic Regression and Decision Trees and Neural Networks for accuracy and generalization capabilities [36]. The combination of SVM and Random Forest proved most effective for university grading prediction, but Logistic Regression served as an efficient baseline comparison for performance evaluation [37].

Multiple studies have evaluated different machine learning models to determine their effectiveness for student performance prediction within the field [38]. It has been observed that deep learning algorithms and ensemble methods achieve higher accuracy and better generalization compared to models like Logistic Regression, Decision Trees, and Random Forests. [36]. It was confirmed that SVM and Random Forest with Neural Networks established themselves as top performers in university-grade prediction, yet Logistic Regression supplied fundamental performance data for comparison objectives [39].

It was argued that many essential research gaps still exist after significant advances in the field. On the other hand, little research has yet been conducted into how soluble models like Neural Networks and can be made interpretable enough to boost educator trust and practical application of the results. Secondly, many studies are based on relatively small or homogeneous datasets and, thus, are limited in their generalizability across educational contexts. Thirdly, there is hardly any work integrating real-time data and adaptive learning systems for dynamic student performance prediction. Finally, there is a need for further investigation into computational requirements and complexity of running more advanced models such as Neural Networks for use in educational settings. In this paper, we attempt to address some of those gaps based on a comparison of Logistic Regression, Decision Trees, Random Forest, SVM, and

Neural Networks on university-level data to offer insights into how the best approaches to predict educational tasks at university level.

3. Methodology

3.1 Dataset Overview

In this study, the dataset source from the UCI Machine Learning Repository is considered a known and widely used resource for machine learning research and experiments.. It includes observations of 4,424 students on 37 different columns that describe attributes in the feature space of students along the lines of their demographic, academic and socioeconomic characteristics.

3.2 Data Preprocessing

3.2.1 Correlation Analysis

A correlation analysis was conducted to identify the relationships between various features in the dataset and their potential influence on student academic performance. The primary objective of this step was to assess the strength and direction of linear relationships between features and to identify multicollinearity issues that might affect the predictive models.
The linear relationship among them was visualized using the Pearson correlation coefficient (r) for all numerical features in the dataset (See. Figure 1). The Seaborn library in Python was used to generate a heatmap of the correlation matrix. The heatmap shows correlation values as a visualization, where positive values are graduated in red and negative values in blue.

Strong correlations ($|r| > 0.7$) between features indicate potential redundancy, leading to multicollinearity in regression models and negatively impacting model performance. Features with weak correlations ($|r| < 0.3$) to the target variable may have limited predictive power.
A focused correlation analysis was then performed on academic performance-related features, including first and second-semester curricular units and their respective grades, evaluations, and approvals, along with the target variable. This focused analysis aimed to identify key predictors of student performance.
The heatmap (Figure 2.) reveals the following key insights:

- Curricular units (evaluations) and (approved) exhibit strong positive correlations with semester grades ($r > 0.8$), indicating that higher evaluation scores and approvals are associated with higher grades.
- The target variable (representing student performance outcomes) shows moderate positive correlations with Curricular units 1st sem (grade) and Curricular units 2nd sem (grade), suggesting that semester grades are significant predictors of overall performance.
- Curricular units without evaluations negatively correlate with semester grades, indicating that missing evaluations may indicate lower academic performance.

This correlation analysis provided critical insights into feature relationships and guided feature selection for subsequent machine learning modeling, ensuring the selection of relevant predictors while mitigating multicollinearity issues.

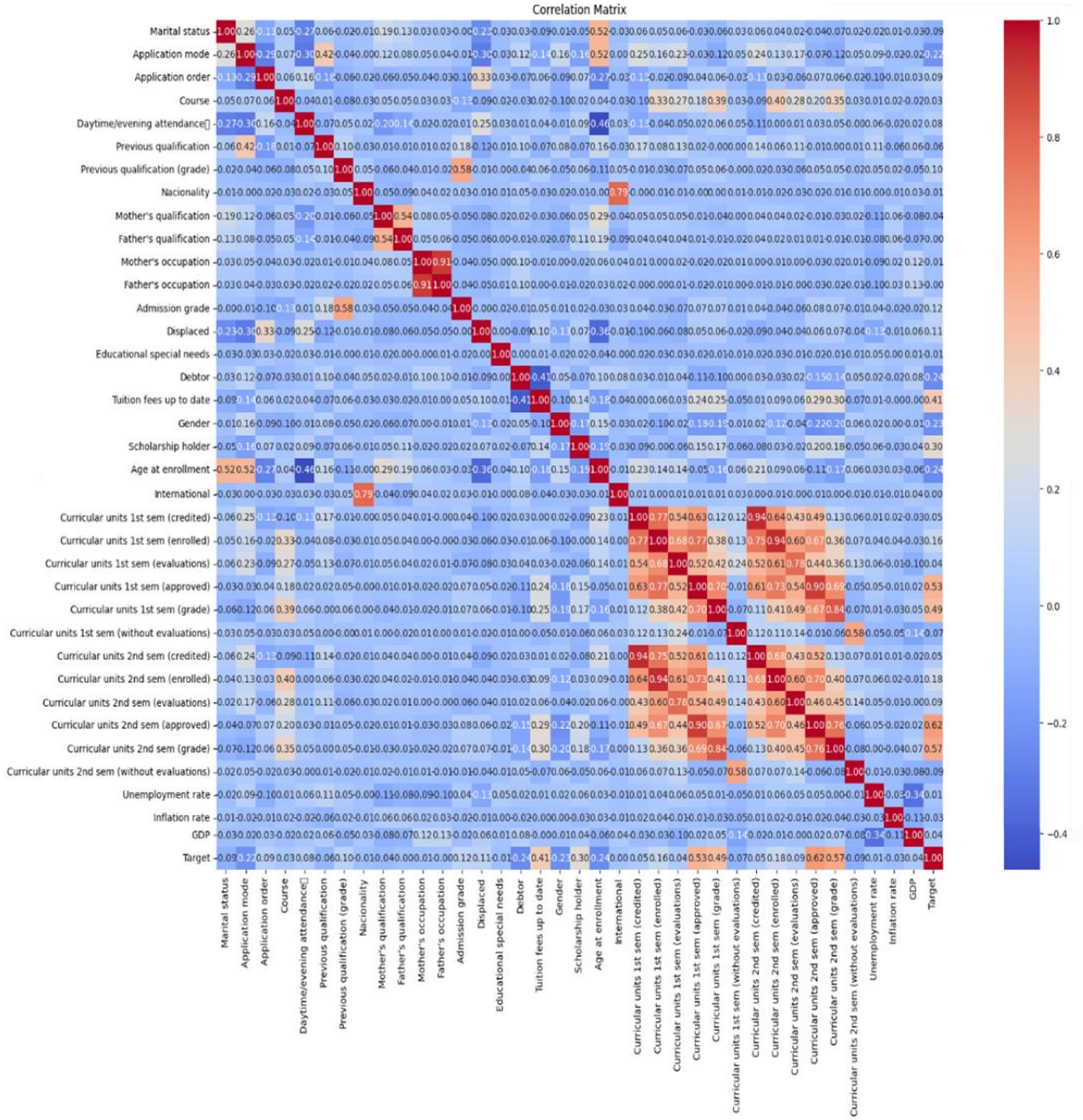


Figure 1: Correlation Matrix for All Features

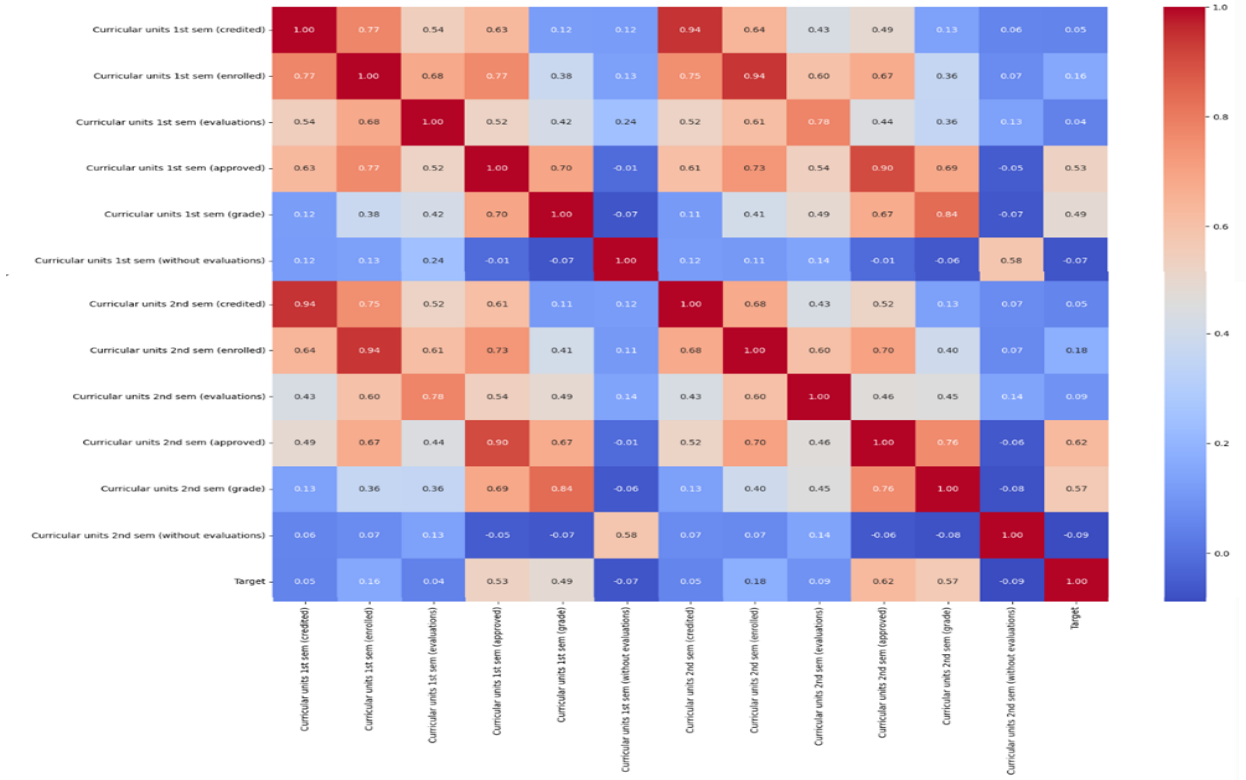


Figure 2: Correlation Matrix for Academic Performance Features

3.2.2 Feature Selection

Feature selection is crucial in machine learning preprocessing as it selects the best pertinent features for predicting the target variable. The dimensionality of the dataset is reduced with the selection of the most informative features, which, in effect, reduces the overfitting, improves model performance, and, in turn, improves the interpretability of the human-created model.

For this study, SelectKBest was utilized to select the top features that will give the best statistical significance to predict the students' performance. The ranking of features came from the f_classif scoring function, which offers the ANOVA F statistic between each feature and the target variable.

The dataset was split into features (X), All columns except the target variable, and target (y), The target variable indicating student performance outcomes.

The SelectKBest method was applied to select the top k features with the highest predictive power. The method starts by ranking the features, and each feature is ranked based on its F-statistic score, which measures the variance between the feature's categories and the target classes. Then, the top features with the highest F-statistic are selected.

As indicated previously, twelve features were identified as the most significant predictors of student performance (See. Figure 2).

3.2.3 Feature Scaling

Feature scaling is especially crucial when the algorithm is sensitive to the absolute value of the input features. This is because of the Support Vector Machine (SVM), K nearest neighbors (KNN), and Logistic Regression. Feature scaling guarantees that the input features have all similar ranges, thus avoiding any preference for features with large numerical ranges.
Standard Scaling was applied to the features to ensure all the selected features were weighted equally to improve the predictive model in this paper. This is very important for models that need the distance between the data points because standard Scaling standardizes the features to have a mean of 0 and a standard deviation of 1.
StandardScaler was selected from the sci-kit-learn library to perform the Scaling. This means that the mean of each feature is subtracted from the mean, and the result is divided by the standard deviation. The formula below is used for standardization.

$$z = \frac{x - \mu}{\sigma}$$

Where:
x *= original feature value.*
μ *= mean of the feature.*
σ *= standard deviation of the feature.*
z *= standardized (normalized) value.*

Finally, scaling was applied to the features chosen according to the feature selection phase for uniformity in the data range. Then, the transformed features were used as input to some machine learning models. Then, the standard scale of all the features is set, so they have a mean of 0 and standard deviation of 1; thus, all the features are on the same scale. Standards of this kind improve optimization algorithms' convergence rate and prevent features of a greater magnitude from dominating the learning process.

3.2.4 Data Splitting

The dataset was divided into training and testing subsets to assess the model's ability to generalize to unseen data. Specifically, 80% of the data was allocated for training the model, while the remaining 20% was reserved for testing. The training set was used to fit the model, and the testing set served to evaluate its performance on data it had not encountered during training.

3.3 Data Analysis Algorithms

Over the past decades, predictive analytics and ML algorithms have become indispensable tools that help replace intuitions and basic statistics with machine learning. For this study, various supervised machine learning algorithms were applied to predict student performance based on the available student academic data. Each algorithm has unique strengths and weaknesses in terms of accuracy, interpretability, and computational efficiency.
The algorithms used in this study include a linear model, Logistic Regression, which is often used as a baseline for classification tasks, non-parametric methods such as Decision Trees, an interpretable yet straightforward method, the assembly of multiple decision trees through Random Forest, which aims at increasing accuracy by aggregating multiple decision trees, and the robust classifier used in high dimensional spaces Support Vector Machines (SVM). Then, these algorithms are selected for various approaches, from interpretable linear models to more complex nonlinear models, so that a complete evaluation of their effectiveness in predicting student performance can be considered. In the next sections,

details of each algorithm, including its theory, implementation, and performance evaluation, will be presented.

3.3.1 Logistic Regression

Logistic Regression is one of the most widely used classification algorithms developed to represent a dependent variable with the help of one or more independent variables using the logistic function. Furthermore, this study uses logistic Regression to predict student performance, which is represented as a target variable (categorical). It was used because of its simplicity, interpretability and effectiveness in binary and multi-classification problems.
In this study, the sci-kit-learn library's Logistic_Regression model was used. A feature scaling was done, and the data was split on the training dataset to train the logistic regression model. The model predicts what observation comes from a given class (in this case, poor, average or high-performance categories for the students). The model parameters used were as follows:

- C = 1.0: Regularization strength, controlling overfitting. A lower value of C would increase regularization, whereas higher values allow for less regularization.
- Solver = 'lbfgs': The optimization algorithm to find the best parameters. The 'bugs' solver is a quasi-Newton method suitable for smaller datasets.
- Max_iter = 100: The maximum number of iterations the algorithm will use to converge to a solution.
- Multi_class = 'auto': Logistic Regression in the multi-class setting uses a one-vs-rest scheme by default when there are more than two classes.

The applied default parameters serve effectiveness and simplicity purposes, but hyperparameter tuning would enhance performance. The logistic regression model generates interpretable outcomes where each coefficient represents the limit between feature variables and target class probability.

3.3.2 Decision Tree

Decision Trees serve as one of the most frequently used classification algorithms in machine learning. The tree structure serves as the data model where nodes at the inner levels contain decision-based features, and leaf nodes provide predictions for class labels. The use of Decision Trees remains beneficial because they connect easily with users while accepting numeric and categorical input data. The research applied a Decision Tree Classifier as a predictive model.
The research implemented Decision_Tree_Classifier to build the model. The initialization of the classifier used random_state=42 as a parameter, which established result reproducibility, including consistent data splits in each execution cycle.
After the model training process, the model produced predictions for both the training and test data sets. Analyses of model fitting performance took place by applying predictions to the testing data set. The model parameters used were as follows:

- Criterion = 'gini': The function employed to assess the quality of a split. The default Gini Impurity quantifies the frequency of incorrect classifications for a randomly selected element.
- Max_depth = None: specifies the maximum depth of the decision tree. When set to None, the tree will expand nodes until all leaves are pure (i.e., contain only one class) or until each leaf has fewer samples than the minimum required to perform a split.

• Min_samples_split = 2: This parameter defines the minimum number of samples necessary to divide an internal node.

• Min_samples_leaf = 1: This parameter specifies the minimum number of samples necessary for a leaf node.

• Random_state = 42: This parameter guarantees the reproducibility of results by establishing a fixed random seed.

This analysis employed default hyperparameters for simplicity; however, further tuning of parameters like max_depth or min_samples_split may enhance model performance by mitigating overfitting. The Decision Tree model exhibits high interpretability, as its structure offers clear insights into decision-making. The tree bifurcates according to various features' significance, allowing for visualization to elucidate the model's predictive mechanisms.

3.3.3 Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees during the training process and predicts the output class based on the majority vote (mode) of the individual trees in classification tasks. It is well-regarded for its high accuracy, resistance to overfitting, and capability to handle large datasets with complex feature interactions. Due to its effectiveness and flexibility, Random Forest is widely applicable across various classification tasks.

The model initially involved the following essential parameters: The random_state attribute set to value 42 allows the model to split data identically during each run by locking down the random seed.

• n_estimators=100: Specifies the number of decision trees in the forest. The model used a total of one hundred trees for its construction. The model quality improves when adding more estimators, yet additional benefits tend to plateau at a specific point. We used predictions to assess the model's operation on training data and measure its capacity to work with new, unobserved information.

3.3.4 SVM

Users use SVM as a robust supervised machine learning algorithm dedicated to solving classification problems. The algorithm identifies the optimal hyperplane, creating the biggest distance between support vector points representing different class categories. SVM operates effectively in high-dimensional spaces and data sets, which are non-linearly separable but become most powerful when it uses different kernel functions. The research implemented a Support Vector Machine with a linear kernel to forecast student academic outcomes depending on multiple school characteristics and individual information.

User implementation of SVM classification occurred through the SVC (Support Vector Classification) class provided within the sci-kit-learn library. The model parameters used were as follows:

• kernel='linear': The linear kernel was chosen because it works well for linearly separable data. It transforms the data into a higher-dimensional space where a linear separation is possible.

• random_state=42: This ensures the results are reproducible by fixing the random seed, leading to the same data split every time the model is run.

Predictions were made on the training set (X_train, y_train) on both the training and testing data.

The predictions on the training data help assess how well the model learned from the training set, while predictions on the testing data evaluate the model’s generalizability to unseen data.

3.3.5 Neural Network

Neural Networks' sophisticated classification algorithm operates based on biological neural systems to solve complex problems during deep learning processes. Data passes through neural connectivity layers where nonlinear activation functions transform processed information to function. The distinctive design structure of Neural Networks helps identify complex education patterns, thus making them ideal for evaluating student academic success based on various educational and demographic variables. The Neural Network classification became possible through the MLPClassifier (Multi-Layer Perceptron) class from sci-kit-learn Library. The model parameters used were as follows:

- hidden_layer_sizes=(100): Specifies the architecture with one hidden layer containing 100 neurons, providing sufficient complexity to capture nonlinear relationships in the educational data.
- max_iter=500: Sets the maximum number of iterations for the solver to converge, ensuring adequate training time for complex pattern recognition.
- random_state=42: Ensures reproducibility by fixing the random seed, maintaining consistent results across multiple runs.

Both training and testing datasets received predictions after processing the training through the model training procedure. The predictions resulting from training data measure the learning capabilities of the model, whereas predictions generated from testing data determine its performance on new, unfamiliar data.

4. Results and Discussion

4.1 Logistic Regression Analysis

As shown in Table 1, Testing data from the Logistic Regression model delivered 69.27% accuracy and 78.64% F1-score measurement. As illustrated in Figure 3, a model analysis using the confusion matrix showed that 212 cases and 394 cases in the classes matched the correct predictions. In comparison, at least seven instances experienced misclassification based on nearest class assessments. Two-thirds of positive predictions from the model were accurate according to precision measurement at 67.5%, but specificity evaluation at 59.3% indicates errors in its negative predictions.

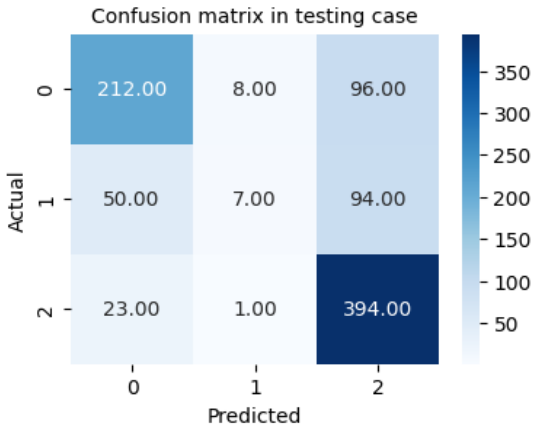


Figure 3: Confusion matrix of Logistic Regression

Table 1. Performance Measures for Logistic Regression

Accuracy%	Precision%	Specifity%	F1 Score%
69.27	0.674658	0.593148	0.786427

4.2 Decision Tree Analysis

Figure 4 reveals important aspects of the Decision Tree model's classification outcomes in the confusion matrix. The model presented the highest accuracy in classifying instances to their respective groups, where 206 class 0 and 303 class 2 labels matched correctly but produced only 46 correct predictions for class 1. The Decision Tree model performs poorly when classifying instances from class 1 since it consistently misinterpreted 49 cases as class 0 while 56 cases belonged to class 2. The model incorrectly classified 48 class 0 instances into class 1 and 62 more class 0 instances into class 2. The metrics presented in Table 2, taken from testing showcase moderate performances through accuracy at 62.71%, precision at 71.97%, specificity at 74.73% and an F1 score at 72.22%.

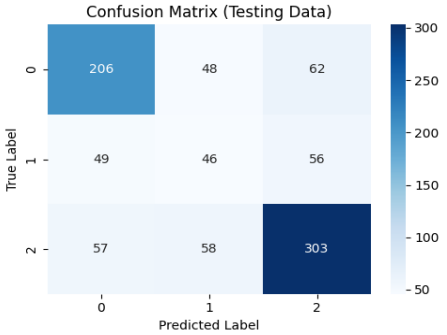


Figure 4: Confusion matrix of Decision Tree

Table 2. Performance Measures for Decision Tree

Accuracy%	Precision%	Specifity%	F1 Score%
62.71%	71.97%	74.73%	72.22%

4.3 Random Forest Analysis

The Random Forest model excels beyond the Decision Tree in accuracy terms as represented in Figure 5 by its confusion matrix, which shows a significant increase in class 2 prediction success, 340 from 303. The model correctly classifies 207 cases of class 0 but remains limited to class 1, achieving only 36 correct identifications. The analysis shows equal misclassification frequencies between class 0 and class 2 at 44 instances and 44 instances, while class 0 experiences 44 misclassifications to class 1 and 65 to class 2. Table 3 demonstrates the Random Forest moderate improvements in performance through an accuracy rate of 65.87% together with a precision of 71.42%, specificity of 70.87%, and F1 score of 76.06% but shows difficulties in separating class 1 from other classifications.

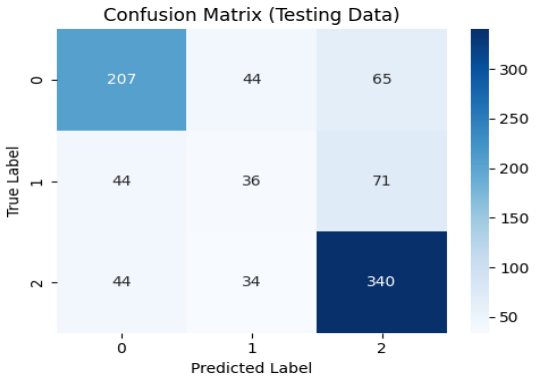


Figure 5: Confusion matrix of Random Forest Analysis

Table 3. Performance Measures for Random Forest Analysis

Accuracy%	Precision%	Specifity%	F1 Score%
65.87%	71.42%	70.87%	76.06%

4.4 SVM Analysis

The performance characteristics of the SVM model differ from that of the decision tree and random forest models. While delivering specific advantages, it also encounters matching trade-offs. Class 2 receives the best classification from the confusion matrix presented in Figure 6, by achieving 394 correct predictions, which surpasses both previous models, while class 0 performed with 204 correct predictions. This model demonstrates the poorest effectiveness among all three in classifying class 1, achieving 15 correct predictions as its minimum total. The model strongly prefers class 2 since it misclassified 95 instances from class 0 and 94 cases from class 1. As shown in Table 4, the SVM produces an improved accuracy rate of 69.26% compared to prior models while demonstrating lower specificity at 59.52% because its classification performance is uneven across all classes, especially for class 1.

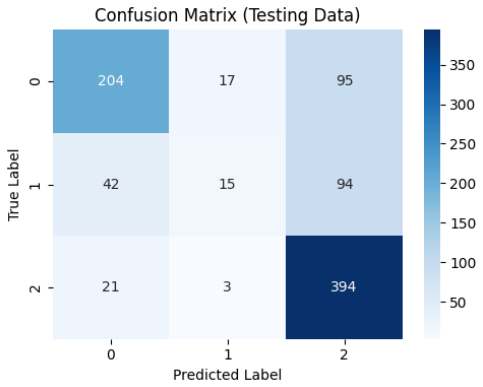


Figure 6: Confusion matrix of SVM

Table 4. Performance measures for SVM

Accuracy%	Precision%	Specifity%	F1 Score%
69.26%	67.58%	59.52%	78.72%

4.5 Neural Network Analysis

All evaluated models demonstrate that the Neural Network model delivers the maximum overall performance through multiple strategic improvements. As presented in Table 5, out of all the assessed models, this system shows the highest accuracy rate, 69.83%, and the F1 score, 80.29%. Figure 7 shows that a total of 381 correct predictions for class 2 ranks second after SVM's outcome, while achieving 209 correct predictions for class 0 places second to XGBoost's 216 predictions in the confusion matrix. The neural network model classifies class 1 errors at a rate of 28 cases, better than SVM (15) yet worse than Decision Tree (46), with 29 correct predictions. The model displays continuous misjudgment toward class 2 because it erroneously classifies 75 cases from class 0 and class 1 as class 2. The Neural Network offers the most even distribution of errors between classes than other models. Although its specificity stands at 67.88%, it maintains a superior accuracy of 71.75% through balanced error distribution, thus enabling the best trade-off between precision and generalization in this study.

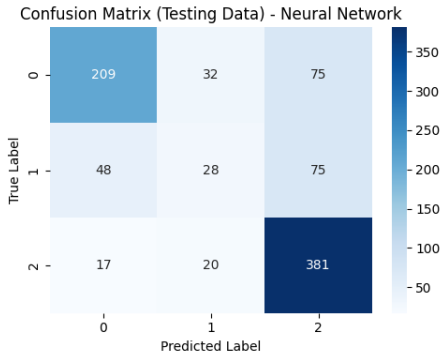


Figure 7: Confusion matrix of Neural Network

Table 5. Performance Measures for Neural Network

Accuracy%	Precision%	Specifity%	F1 Score%
69.83%	71.75%	67.88%	80.29%

4.6 Discussion

The extensive performance evaluation of machine learning models produced results that validate research findings from the literature review. Results demonstrate that Neural Networks delivered the best performance by achieving 69.83% accuracy and 80.29% F1 score since they are known to identify complex nonlinear educational patterns, according to [32]. Models demonstrate higher performing accuracy metrics because they recognize patterns of student behavior that traditional models cannot detect, as explained by [33].

SVM achieved identical accuracy numbers to XGBoost at 69.26% yet reported lower specificity at 59.52% based on the sophisticated mathematical principles described by [29] in their work on education data analyses. The model achieved 394 correct classified instances for class 2, which supports its powerful pattern detection ability in processing complex education datasets per [30].

Random Forest produced 65.87% accuracy results but displayed balanced evaluation across all measures, with an F1 score of 76.06%. [24] documented that Random Forest exhibits a strong capability to process intricate dataset relations to decrease overfitting potential, according to their study. The results demonstrate the consistent predictive ability of Random Forest as described in [26].

The Decision Tree model achieved an accuracy rate of 62.71%, which placed it last among all models, though it continued showing reasonable performance for other metrics. The findings support [19], who

noted that decision trees provide simple interpretability features along with visualization capabilities but typically yield inferior predictive results compared to advanced models. The visual quality of decision trees provides substantial value to educational practitioners who must interpret the decision-making process, according to [22].

The research findings back up existing scientific understanding, regarding Neural Networks surpassing other methods in detecting complex educational data patterns. The results from the study confirm [35] findings, which suggest ensemble methods and complex models achieve better results than basic Decision Trees and Logistic Regression models when evaluating accuracy and generalization skills.

5. Conclusion

An extensive study of machine learning algorithms for student academic prediction has produced multiple important findings that enhance educational data mining and analytics research. Five distinct machine learning models, including Neural Networks, Support Vector Machines, Random Forest, Decision Trees and Logistic Regression, show more sophisticated algorithms produce better predictions for educational purposes. Research showed that Neural Networks delivered the best results by achieving 69.83% accuracy while obtaining an 80.29% F1 score, which proves that complex nonlinear approaches excel in capturing students' academic performance complexity. The results support educational institutions adopting better analytical systems to detect complex linkages between academic and demographic aspects and student achievement behavior.

Since decision trees provide educational value with simple structures, they remain interpretable despite their accuracy of 62.71%. Predominant organizations need to find a proper equilibrium between sophisticated model achievements and straightforward interpretability when they adopt predictive analytics solutions.

Future research should concentrate on multiple priority areas. Real-time student datasets integrated into predictive models will boost their analytical capabilities, resulting in immediate intervention opportunities. Numerous feature selection techniques should be studied to achieve better model performance results. Research about explainable complex models used in education, such as Neural Networks, may create opportunities to unite predictive strength with understandable outputs.

The research limitations from a specific institutional context and a fixed variable set during the study create potential spaces for validation across various educational institutions and student demographics. The study would benefit from future examinations of how various additional factors, including student engagement statistics, extracurricular actions, and psychological elements, affect accuracy levels of predictive outcomes.

This research shows robust evidence that advanced machine learning methods should be used for educational purposes, yet simpler interpretable models still have their place. The study expands educational data mining literature while providing operational recommendations for institutions that want to adopt data-based approaches to support student achievement prediction.

References

[1] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, “Real-time attention monitoring system for classroom: A deep learning approach for student’s behavior recognition,” Big Data Cogn. Comput., vol. 7, no. 1, p. 48, 2023.

[2] R. McHaney, The New Digital Shoreline: How Web 2.0 and Millennials Are Revolutionizing Higher Education. Taylor & Francis, 2023.
[3] M. A. Hashim, I. Tlemsani, and R. Matthews, “Higher education strategy in digital transformation,” Educ. Inf. Technol., vol. 27, no. 3, pp. 3171–3195, 2022.
[4] V. Pendakur, Ed., Closing the Opportunity Gap: Identity-Conscious Strategies for Retention and Student Success. Taylor & Francis, 2023.
[5] X. Bai et al., “Educational big data: Predictions, applications and challenges,” Big Data Res., vol. 26, p. 100270, 2021.
[6] A. Mahboubi et al., “Evolving techniques in cyber threat hunting: A systematic review,” J. Netw. Comput. Appl., p. 104004, 2024.
[7] W. Seo and Y. Bu, “Transforming role classification in scientific teams using LLMs and advanced predictive analytics,” arXiv preprint arXiv:2501.07267, 2025.
[8] M. M. Asad and A. Qureshi, “Impact of technology-based collaborative learning on students’ competency-based education: Insights from the higher education institution of Pakistan,” High. Educ. Skills Work-Based Learn., 2025.
[9] M. A. Alrowaily et al., “Modeling and analysis of proof-based strategies for distributed consensus in blockchain-based peer-to-peer networks,” Sustainability, vol. 15, no. 2, p. 1478, 2023.
[10] G. Oliveira, J. Grenha Teixeira, A. Torres, and C. Morais, “An exploratory study on the emergency remote education experience of higher education students and teachers during the COVID-19 pandemic,” Br. J. Educ. Technol., vol. 52, no. 4, pp. 1357–1376, 2021.
[11] A. Malik et al., “Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model,” Array, vol. 19, p. 100303, 2023.
[12] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, “A systematic review of deep learning approaches to educational data mining,” Complexity, vol. 2019, no. 1, p. 1306039, 2019.
[13] H. Almaghrabi, B. Soh, A. Li, and I. Alsolbi, “SoK: The impact of educational data mining on organisational administration,” Information, vol. 15, no. 11, p. 738, 2024.
[14] V. Onker et al., “Harnessing machine learning for academic insight: A study of educational performance in Bhopal, India,” Educ. Inf. Technol., pp. 1–40, 2025.
[15] A. Ahmad, S. Ray, M. T. Khan, and A. Nawaz, “Student performance prediction with decision tree ensembles and feature selection techniques,” J. Inf. Knowl. Manag., p. 2550016, 2025.
[16] Y. Salunke et al., “Fraud detection: A hybrid approach with logistic regression, decision tree, and random forest,” Cureus, vol. 17, no. 1, 2025.
[17] A. Leiva-Araos, C. Contreras, H. Kaushal, and Z. Prodanoff, “Predictive optimization of patient no-show management in primary healthcare using machine learning,” J. Med. Syst., vol. 49, no. 1, p. 7, 2025.
[18] S. Alturki, I. Hulpuș, and H. Stuckenschmidt, “Predicting academic outcomes: A survey from 2007 till 2018,” Technol. Knowl. Learn., vol. 27, no. 1, pp. 275–307, 2022.
[19] J. Fan, “A big data and neural networks driven approach to design students management system,” Soft Comput., vol. 28, no. 2, pp. 1255–1276, 2024.
[20] N. I. Elzayat et al., “A comparative study for survival prediction of NKI data using statistical and machine-learning approaches,” in 2024 6th Novel Intelligent and Leading Emerging Sciences Conf. (NILES), pp. 269–273, Oct. 2024.
[21] M. Banerjee, E. Reynolds, H. B. Andersson, and B. K. Nallamothu, “Tree-based analysis: A practical approach to create clinical decision-making tools,” Circ. Cardiovasc. Qual. Outcomes, vol. 12, no. 5, p. e004879, 2019.

[22] M. El Jihaoui, O. E. K. Abra, and K. Mansouri, “Factors affecting student academic performance: A combined factor analysis of mixed data and multiple linear regression analysis,” IEEE Access, 2025.
[23] X. Wan, J. Zeng, and L. Zhang, “Predicting online shopping addiction: A decision tree model analysis,” Front. Psychol., vol. 15, p. 1462376, 2025.
[24] J. F. Hair Jr, L. P. Fávero, W. T. Junior, and A. Duarte, “Deterministic and stochastic machine learning classification models: A comparative study applied to companies’ capital structures,” Mathematics, vol. 13, no. 3, p. 411, 2025.
[25] X. Bai, L. Zhang, Y. Feng, H. Yan, and Q. Mi, “Multivariate temperature prediction model based on CNN-BiLSTM and RandomForest,” J. Supercomput., vol. 81, no. 1, p. 162, 2025.
[26] Z. Khoudi, N. Hafidi, M. Nachaoui, and S. Lyaqini, “New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments,” SN Comput. Sci., vol. 6, no. 2, p. 139, 2025.
[27] S. Zhao, D. Zhou, H. Wang, D. Chen, and L. Yu, “Enhancing student academic success prediction through ensemble learning and image-based behavioral data transformation,” Appl. Sci., vol. 15, no. 3, p. 1231, 2025.
[28] S. Alghamdi, B. Soh, and A. Li, “A comprehensive review of dropout prediction methods based on multivariate analysed features of MOOC platforms,” Multimodal Technol. Interact., vol. 9, no. 1, p. 3, 2025.
[29] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, “Exploring kernel machines and support vector machines: Principles, techniques, and future directions,” Mathematics, vol. 12, no. 24, 2024.
[30] A. Bassi, A. A. Mir, B. Kumar, and M. Patel, “A comprehensive study of various regressions and deep learning approaches for the prediction of friction factor in mobile bed channels,” J. Hydroinform., vol. 25, no. 6, pp. 2500–2521, 2023.
[31] S. HAKKAL and A. A. LAHCEN, “XGBoost to enhance learner performance prediction,” Comput. Educ.: Artif. Intell., p. 100254, 2024.
[32] M. A. Shyaa et al., “Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems,” Eng. Appl. Artif. Intell., vol. 137, p. 109143, 2024.
[33] G. Alhussein, M. Alkhodari, A. H. Khandoker, and L. J. Hadjileontiadis, “Novel speech-based emotion climate recognition in peers’ conversations incorporating affect dynamics and temporal convolutional neural networks,” IEEE Access, 2025.
[34] H. Pallathadka et al., “Classification and prediction of student performance data using various machine learning algorithms,” Mater. Today: Proc., vol. 80, pp. 3782–3785, 2023.
[35] M. M. Rahaman, S. Rani, M. R. Islam, and M. M. R. Bhuiyan, “Machine learning in business analytics: Advancing statistical methods for data-driven innovation,” J. Comput. Sci. Technol. Stud., vol. 5, no. 3, pp. 104–111, 2023.
[36] N. Rane, S. P. Choudhary, and J. Rane, “Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions,” Stud. Med. Health Sci., vol. 1, no. 2, pp. 18–41, 2024.
[37] G. Latif, S. E. Abdelhamid, K. S. Fawagreh, G. B. Brahim, and R. Alghazo, “Machine learning in higher education: Students’ performance assessment considering online activity logs,” IEEE Access, vol. 11, pp. 69586–69600, 2023.
[38] S. Batool et al., “Educational data mining to predict students' academic performance: A survey study,” Educ. Inf. Technol., vol. 28, no. 1, pp. 905–971, 2023.

[39] A. Kye, "Comparative analysis of classification performance for US college enrollment predictive modeling using four machine learning algorithms (logistic regression, decision tree, support vector machine, artificial neural network)," Ph.D. dissertation, Loyola Univ. Chicago, 2023.