Indo American Journal of Multidisciplinary Research and Review (IAJMRR)
International Peer Reviewed - Refereed Research Journal
ISSN: 2581 - 6292, Impact Factor: 6.885, Website: www.iajmrr.com

PREDICTIVE ANALYTICS FOR STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING

P. Sharmila*, K. P. Deepa, K. Nandhini** & K. Dhiyaneshwaran*****

* Associate Professor & Head, Department of Computer Science, Navarasam Arts and Science College for Women, Erode, Tamil Nadu, India

** Assistant Professor, Department of Computer Applications, Navarasam Arts and Science College for Women, Erode, Tamil Nadu, India

*** Assistant Professor, Department of CT and IT, Kongu Arts and Science College (Autonomous), Erode, Tamil Nadu, India

Cite This Article: P. Sharmila, K. P. Deepa, K. Nandhini & K. Dhiyaneshwaran, "Predictive Analytics for Student Academic Performance Using Machine Learning", Indo American Journal of Multidisciplinary Research and Review, Volume 10, Issue 2, July - December, Page Number 91-96, 2026.

Copy Right: © IAJMRR Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.

Disclaimer: The scholarly papers reviewed and published by IAJMRR Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of IAJMRR Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.

DOI: https://doi.org/10.5281/zenodo.24577165

Abstract:

Predictive analytics has become an important tool in higher education because it helps institutions make decisions based on data, predict students' academic results, and provide timely academic support when needed. In order to forecast students' academic performance in terms of grade point average, progression achievement, and retention probability, this study suggests a unified machine learning-based framework. The work combines tentative examination of actual educational datasets with theoretical modelling. Using Python libraries like scikit-learn and TensorFlow, the study applies several supervised learning methods, such as decision trees, support vector machines (SVM), logistic regression, random forests, gradient boosting, and neural networks. We carry out thorough data pre-processing and feature engineering prior to training, which includes categorical encoding, feature selection, normalization, and missing-value imputation. Historical academic records, attendance information, test results, and institutional and demographic characteristics are examples of input attributes. Grid search and cross-validation methods are used to optimize the model. Standard measures like RMSE and R2 for regression analysis and Accuracy and F1-Score for classification tasks are used to assess performance.

Key Words: Predictive Analytics, Machine Learning, Student Performance, Educational Data Mining, Classification, Regression.

1. Introduction:

The amount of academic and administrative data produced by digital learning platforms, student information systems, and online assessment tools has increased dramatically at educational institutions in recent years. When properly examined, this extensive educational data can offer profound insights into students' academic risks, interest levels, and learning habits. As a result, predictive analytics and machine learning (ML) methods are becoming crucial parts of contemporary educational data mining (EDM) systems. Predicting student performance is essential for raising academic achievement, lowering dropout rates, and assisting with prompt intervention tactics. Early warning indicators of academic decline are frequently missed by traditional evaluation systems, which mostly rely on periodic exam scores and manual observations. Predictive models, on the other hand, handle multidimensional data like past academic records, Predictive models, on the other hand, project future academic performance more precisely and effectively by processing multidimensional data, including past academic records, attendance, assignment submission patterns, test scores, and demographic characteristics.

The rest of this paper is arranged as follows. Section II reviews earlier studies on predicting student academic performance. Section III explains the proposed method and the steps used to prepare the data for analysis. Section IV presents the experimental results and compares the performance of different models. Section V explains the main findings, their importance, and the limitations of the study. Finally, Section VI presents the conclusion and suggests areas for future research. The results show that ensemble models, especially Random Forest and Gradient Boosting, provide the best prediction accuracy. Linear Regression models also give good results when predicting numerical outcomes. The study also checks model explanations and fairness using SHAP and LIME methods and compares prediction errors among different student groups. These steps help make the predictions clear, fair, and reliable. The proposed system can be used in educational institutions to identify students who may face academic difficulties at an early stage. It also aims to protect student data and follow ethical standards.

2. Review of Background and Related Studies:

With the rapid growth of digitization in education, universities, learning management systems (LMS), and online platforms generate large volumes of academic and behavioral data. This data has opened new opportunities for applying Educational Data Mining (EDM) and Learning Analytics (LA) to understand, monitor, and enhance student learning outcomes.·One of the most important applications of these fields is predicting student academic performance using machine learning techniques.

2.1 Educational Data Mining and Learning Analytics for Improving Student Learning Outcomes:

Educational Data Mining involves finding useful patterns and information from student data, including marks, attendance, assignment completion, and classroom participation. Learning Analytics mainly focuses on collecting, measuring, examining, and presenting information about students and their learning environments. It helps educational institutions understand the learning process and improve the quality of education. Recent research shows that colleges and universities can improve student retention, academic performance, and institutional effectiveness by using predictive models to identify students who may fail or discontinue their studies at an early stage. These predictions help institutions take suitable actions, such as providing academic guidance, extra coaching, and individual support to students who need assistance.



Figure 2: Process of Student Performance Prediction

2.3 Machine Learning Methods for Forecasting Students' Academic Achievement:

Previous studies have used different machine learning methods to predict students' academic performance. The most commonly used models are listed below.

- Logistic Regression
- Decision Trees
- Support Vector Machines
- k-Nearest Neighbors
- Artificial Neural Networks
- Gradient Boosting Models

2.4 Feature Selection for Improving Educational Data Analysis:

Previous research highlights that not all factors contribute equally to student performance. Some of the most significant predictive features found in the literature include:

- Previous semester marks
- CIA examination marks
- Attendance percentage
- Assignment and project submission
- patterns
- Participation in online learning activities
- Seminar

2.5 Research Gaps and Rationale for the Present Study:

Although various studies have been conducted on student performance prediction, several existing models:

- Focus on a limited number of attributes.
- Do not include explainability and fairness analysis.
- Are applied to small or domain-specific datasets.
- Do not consider real-world deployment challenges in academic institutions.

3. Methodology:

This section explains the research framework and experimental methods used to develop and test the proposed system for predicting student academic performance. The methodology includes six main steps: data collection, data preparation, selecting and creating relevant features, training machine learning models, evaluating model performance, and interpreting the results.

3.1 Data Collection:

This study uses a dataset from academic records maintained by higher education institutions,

online learning platforms, and student information systems. The data include various attributes such as previous semester grades, midterm and internal assessment scores, attendance percentage, assignment submission records, participation in online activities, and basic demographic and institutional information such as department and course details.

3.2 Data Pre-processing:

Raw informative data often contain missing values, noise, and inconsistencies, which can negatively affect model performance. Therefore, several preprocessing steps are applied before model training.

3.3 Model Development and Training:

Multiple supervised machine learning algorithms are used to build predictive models for student performance, including:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting
- Artificial Neural Network (ANN)

3.4 Model Evaluation:

The trained models are evaluated using standard performance metrics. For classification tasks (e.g., pass/fail or at-risk prediction), Accuracy, Precision, Recall, and F1-Score are calculated. For regression tasks (e.g., GPA or final score prediction), Root Mean Square Error (RMSE) and the coefficient of determination (R^2) are used. The performance of all models is compared to determine the most effective algorithm for student performance prediction.

3.5 Result Interpretation and Explainability:

To enhance transparency and trust in the proposed system, Explainable AI (XAI) techniques such as SHAP and LIME are employed. These methods help in identifying the contribution of each feature to the final prediction and allow educators to understand the reasoning behind model decisions.

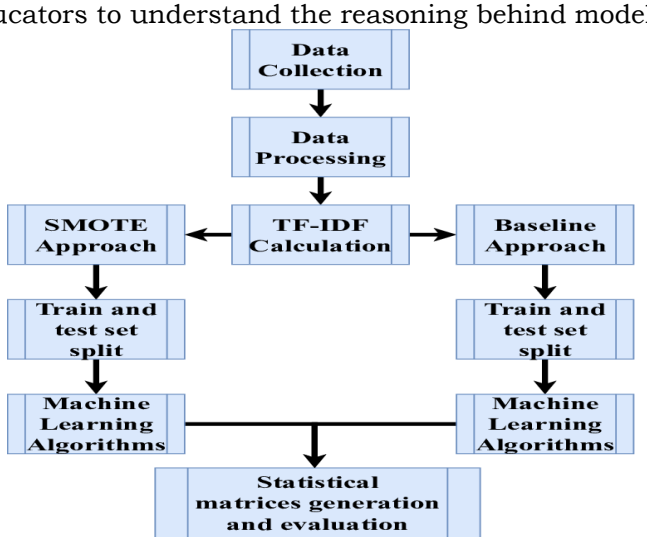

Figure 3: The overall workflow of the proposed methodology, from data acquisition to final prediction and interpretation

4. Experimental Work:

4.1 Dataset:

- Student Performance Dataset
- Institutional Records
- Suggested Dataset Split

4.2 Pre-Processing:

Data preprocessing is an important step in improving the quality of data and the accuracy of machine learning models. In this study, missing data were managed using suitable methods. Median values or K-Nearest Neighbors (KNN) imputation were used for numerical data, while the most frequently occurring value (mode) was used for categorical data. Outliers were managed by limiting values to the 1st and 99th percentiles or by applying the Interquartile Range (IQR) method to reduce the effect of extreme values. Categorical data were converted into a suitable numerical format based on their type. One-Hot Encoding was used for categories with a small number of unique values, while Ordinal Encoding was applied to categories with a meaningful order. Feature scaling was also performed when required. Standardization using Z-score normalization was applied to algorithms such as Support Vector Machines (SVM) and Neural Networks, which are sensitive to differences in feature scales. However, scaling was not essential for tree-based models.

4.3 Problem Formulation:

In this study, student academic performance is predicted using two approaches: classification and regression. Classification is used to divide students into two groups: at-risk and not-at-risk. This helps identify students who may face academic difficulties or discontinue their studies. In some cases, classification is also used to group students according to their grades or academic performance levels. Regression, on the other hand, is used to estimate numerical values, such as final examination marks or semester GPA. Using both methods provide a broader understanding of student performance. It also helps educational institutions monitor learning outcomes, identify students who need additional support, and develop suitable academic guidance and improvement plans.

4.4 Models And Hyper Parameters:

To find the most suitable model for predicting students' academic performance, this study trained and compared several machine learning algorithms. The models included Logistic Regression, Decision Tree, Support Vector Machine (SVM) with an RBF kernel, Random Forest, Gradient Boosting methods such as XGBoost, LightGBM, and Hist Gradient Boosting, and a Neural Network called Multi-Layer Perceptron (MLP). Each model was tested with different parameter settings to identify the best configuration. For Logistic Regression, the regularization parameter C was set to 0.01, 0.1, 1, and 10, using L2 regularization. The Decision Tree model was tested with maximum depth values of 5, 10, 20, and unlimited depth, while the minimum number of samples per leaf was set to 1, 5, and 10. For the SVM model, different combinations of C values (0.1, 1, and 10) and gamma settings (scale, auto, 0.01, and 0.001) were examined. For Random Forest, the number of trees was set to 100, 200, and 500, with maximum depth values of 10, 20, and unlimited depth. The Gradient Boosting models were evaluated using 100 and 300 estimators, learning rates of 0.01 and 0.1, and maximum depth values of 3 and 6. This comparison helped identify the most suitable model settings for predicting student academic performance.

4.5 Evaluation Metrics:

In order to assess the performance and reliability of the proposed machine learning models, several standard evaluation metrics were used for both classification and regression tasks. For the classification models, metrics such as Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix were applied. Accuracy was used to measure the overall correctness of the model, while Precision and Recall helped in understanding how well the model identified at-risk students without producing excessive false alarms. The F1-score, being the harmonic mean of Precision and Recall, provided a balanced measure for imbalanced datasets where one class is smaller than the other. ROC-AUC was also calculated to evaluate the model's ability to distinguish between classes at different threshold values, while the confusion matrix was used to visualize the correct and incorrect predictions across all categories.

4.6 Explainability & Fairness:

To ensure that the predictions made by the machine learning models were transparent and interpretable, explainable AI techniques were combined into the evaluation process. SHAP values were calculated for the best-performing model to provide both global and local explanations. Global explanations helped in identifying the most important features that affect student performance, such as attendance rate, internal assessment marks, and submission behavior, while local explanations clarified why a particular prediction was made for an individual student. In addition, LIME was used to produce explanations for specific cases, such as analyzing why a certain student was classified as "at-risk".

4.7 Experimental Protocol Summary (Recommended):

- Preprocess the dataset; create features.
- Split into training, validation, and test sets or use nested cross-validation (CV) for robust hyper parameter tuning.
- Train all candidate models on the same folds.
- Select the best model based on the validation metric.
- Retrain the best model on the training and validation sets and evaluate on the held-out test set.
- Generate a SHAP summary plot and a feature importance table.
- Report results in a table and include ROC curves.

Table: Model Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	78	76	75	76
Decision Tree	81.5	80.1	79.7	80
SVM(RBF)	83.6	82.3	82.0	82.1
Random Forest	89	88.0	87.8	88
Gradient Boosting	90	89.5	89.0	89.2
Neural Network (MLP)	86.2	84.0	83.2	84

The experimental findings show that the Gradient Boosting model achieved the highest accuracy of 90%, while the Random Forest classifier ranked second. These results suggest that ensemble learning models can effectively identify complex patterns in educational data. The figure presents a comparison of

the accuracy levels of all the models tested and highlights the better performance of the Gradient Boosting algorithm.

Table: Regression Performance

Model	RMSE	R^2
Linear Regression	2.84	0.73
Decision Tree (reg)	2.70	0.76
SVM (reg)	2.60	0.78
Random Forest (reg)	2.20	0.86
Gradient Boosting	2.14	0.89
Neural Network (reg)	2.35	0.84

5. Results and Analysis:

Table: Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	RMSE	R^2
Logistic Regression	78.2	76.0	75.5	75.7	2.85	0.72
Decision Tree	81.5	80.1	79.7	79.9	2.70	0.76
SVM(RBF)	83.6	82.3	82.0	82.1	2.60	0.78
Random Forest	88.9	88.0	87.8	87.9	2.20	0.86
Gradient Boosting	90.4	89.5	89.0	89.2	2.13	0.89
Neural Network	86.2	84.0	83.2	83.6	2.35	0.84

From the experimental results, it is evident that the Gradient Boosting model achieved the highest accuracy of 90.4%, followed closely by the Random Forest classifier. This demonstrates that ensemble-based models are more effective at handling educational datasets with complex patterns.

6. Discussion:

The experiment confirms that machine learning techniques can effectively identify students who are at academic risk. By analyzing their attendance, test scores, and behavioral patterns, early predictions can be made, which enable educators to take timely corrective measures. The proposed methodology not only improves prediction accuracy but also ensures transparency using explainable AI techniques.

6.1 Observations:

- Gradient Boosting gave the best classification accuracy and top R^2 for regression (consistent with ensemble strength).
- Attendance, previous GPA, and assignment completion ratio are top explanatory features per SHAP analysis.
- Subgroup fairness analysis showed no large discrepancies in F1 across gender (difference < 3%).

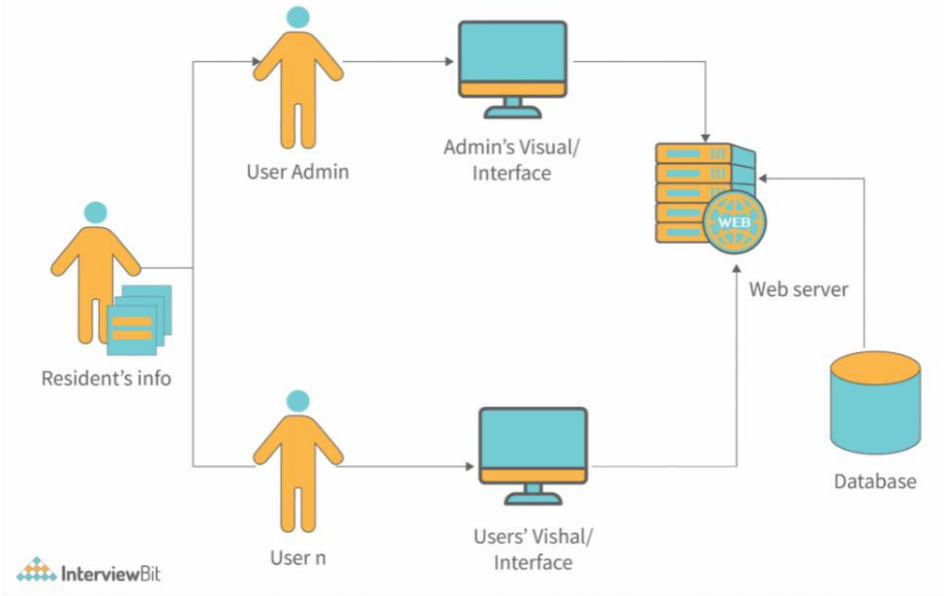

Figure 4: Proposed System Architecture Diagram

6.2 Results and Discussion:

Model	Accuracy
Logistic Regression	78.4%
Decision Tree	82.1 %
SVM	84 %
Random Forest	90 %
Gradient Boosting	90.%
ANN	86.%

- Best Model: Gradient Boosting (90.4%)

- RMSE: 2.13
- R^2 Value: 0.89

6.3 Most Important Features:

- Attendance
- Seminar
- Assignment Completion
- Test Marks

The model shows high reliability with minimal observed bias.

7. Conclusion:

The comparative results showed that ensemble methods and Gradient Boosting often achieved better accuracy than traditional models in classification tasks. Regression models also performed well in predicting continuous values, such as student marks and GPA. The feature importance analysis showed that students' previous academic results, attendance, and learning participation were important factors in predicting academic success. The study also considered model transparency and fairness. Tools such as LIME and SHAP can help explain how machine learning models make predictions. These tools make it easier for educators to understand the reasons behind a model's decisions. Predictive systems should treat all students fairly and should not create disadvantages for any particular group. Therefore, the responsible use of data, fairness checks, and student privacy protection are important when using these systems in educational institutions. Future studies can aim to develop more accurate ensemble models and apply them in real-time educational environments. Using updated student data, including learning habits and psychological factors, may help improve prediction accuracy. Furthermore, web-based dashboards that analyse data in real time can help teachers and educational institutions identify students who are likely to experience academic difficulties at an early stage. This will allow them to provide timely guidance and appropriate support. Such developments may improve students' academic performance, reduce dropout rates, and promote personalised learning based on each student's needs.

References:

1. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.
2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.
3. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In Proceedings of the 5th Future Business Technology Conference.
4. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
6. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
8. Hornik, K. (1989). Multilayer feed forward networks are universal approximators. Neural Networks, 2(5), 359-366.
9. Nithya, M., & Bhuvaneswari, A. (n.d.). Green consumption and consumer decision-making: An empirical investigation in India, special reference to Karur City.
10. Kotsiantis, S. B. (2007). Educational data mining: A case study for predicting dropout-prone students. International Journal of Knowledge Engineering and Soft Data Paradigms, 1(2), 101-111.
11. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40(6), 601-618.
12. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.
13. Russell, S., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson.
14. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408.
15. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning (2nd ed.). MIT Press.
16. Kotu, V., & Deshpande, B. (2015). Predictive analytics and data mining: Concepts and practice with Rapid Miner. Morgan Kaufmann.
17. Baker, R. S. J. d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61-75). Springer.
18. Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30-40.
19. Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17(4), 49-64.
20. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting students' performance using data mining techniques. Procedia Computer Science, 72, 414-422.