© April 2026| IJIRT | Volume 12 Issue 11 | ISSN: 2349-6002

Student Performance Prediction Using Machine Learning

Ashutosh Dev[1], Gaurav Kumar[2], Deepak Gupta[3]
[1]*M.Tech CSE Scholar, IIMT University, Meerut, India*
[2,3]*Assistant Professor, IIMT University, Meerut, India*

***Abstract*—This study investigates the behavioural and academic predictors of academic performance among undergraduate engineering students using supervised machine learning techniques. A dataset comprising demographic profiles, academic records, behavioural patterns, and psychological attributes was analysed. Five machine learning algorithms—Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Naive Bayes—were applied and compared. SHAP (SHapley Additive exPlanations) analysis was employed to identify and rank feature contributions. Random Forest achieved the highest accuracy (91%). Assignment submission regularity, laboratory attendance, and self-study hours emerged as the most influential behavioural predictors. The study recommends targeted behavioural monitoring and early intervention frameworks for engineering programmes.**

***Index Terms*—Engineering Students, Machine Learning, Behavioural Predictors, Feature Importance, SHAP Analysis, Academic Performance, Educational Data Mining**

I. INTRODUCTION

The rapid growth of digital educational data and advancements in computational intelligence have changed how academic institutions monitor and support student learning outcomes. Predicting student academic performance at an early stage allows educators to identify at-risk learners and implement targeted, timely interventions.

Engineering programs are known to be some of the most challenging undergraduate courses. High dropout rates, poor academic performance, and delays in graduation are ongoing issues for engineering institutions worldwide. Unlike other student groups, engineering students face intense laboratory schedules, project-based assessments, and technical coursework that require specific skills and abilities.

Machine learning (ML) provides a strong method for identifying meaningful patterns in large and varied educational datasets. Algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Logistic Regression are commonly used in educational data mining (EDM) to create predictive models [1].

This study addresses the gap by using several supervised machine learning algorithms to an engineering student dataset and employing SHAP analysis to quantify and rank the contribution of each feature to academic performance predictions. The goal is to produce useful, clear insights that can guide targeted institutional interventions.

II. LITERATURE REVIEW

A significant amount of machine learning research has been conducted to predicting student academic outcomes. Yakubu and Abubakar [2] demonstrated that Random Forest was consistently more effective than Logistic Regression and Naive Bayes classifiers. Dabhade et al. [3] confirmed the dominance of ensemble methods across multiple educational datasets.

Orji and Vassileva [4] showed that models incorporate student motivation and study strategies significantly outperformed those relying solely on academic records. Nayak et al. [5] further established that study engagement variables were among the most influential features in predicting performance outcomes.

Rastrollo-Guerrero et al. [6] noted a significant gap in domain-specific research targeting engineering students, observing that most published models were developed using heterogeneous student populations. Balcioglu and Artar [7] advocated for the integration of interpretability tools such as SHAP to bridge the gap between predictive modelling and practical educational decision-making.

Rajendran et al. [8] found that behavioural engagement metrics were stronger predictors than socioeconomic variables among school students. Yagci [9] employed educational data mining for academic performance prediction, further validating the utility of machine learning in diverse educational contexts.

III. RESEARCH GAPS

Despite substantial progress, the following gaps motivated the present study:

- Absence of engineering-specific predictive models focused on behavioural features.
- Limited use of explainability tools such as SHAP in educational ML research.
- The patterns in laboratory attendance and assignments are not well-represented as primary predictors.
- Lack of domain-specific intervention recommendations for engineering programmes.

IV. RESEARCH OBJECTIVES

The present study is guided by the following objectives:

- To identify and rank behavioural and academic predictors of engineering student performance using machine learning.
- To compare predictive accuracy of Random Forest, Decision Tree, SVM, KNN, and Naive Bayes.
- To apply SHAP analysis for transparent feature importance quantification.
- To develop evidence-based recommendations for early academic intervention.

V. METHODOLOGY

A. Dataset and Feature Variables

The dataset was compiled from undergraduate engineering students and encompasses five categories: demographic features (age, gender), academic variables (prior grades, GPA, backlogs), behavioural indicators (lab attendance, study hours, assignment submission rate), psychological attributes (motivation, stress), and engagement metrics (project participation, peer study groups).

B. Data Preprocessing

Preprocessing included: (i) missing value imputation using mean/mode substitution; (ii) outlier treatment using the IQR method; (iii) min-max normalisation; and (iv) one-hot encoding of categorical variables. The target variable was binary: Pass (CGPA $\geq$ 5.0) and At-Risk (CGPA < 5.0).

C. Feature Selection

Recursive Feature Elimination (RFE) was applied to identify the most informative feature subset, reducing dimensionality and overfitting risk.

D. Model Training

The dataset was split into training (70%) and testing (30%) sets using stratified sampling. Five supervised classifiers were implemented: Random Forest, Decision Tree, SVM, KNN, and Naive Bayes. Models were evaluated using accuracy, precision, recall, F1-score, and AUC-ROC.

E. SHAP Analysis

SHAP values were computed for the best-performing model to quantify each feature's marginal contribution to individual predictions, enabling both global and local model interpretation.

VI. RESULTS AND ANALYSIS

A. Model Performance Comparison

Table I presents the comparative performance of the five classifiers evaluated on the engineering student dataset.

TABLE I. CLASSIFICATION PERFORMANCE OF ML MODELS

Algorithm	Acc.	Prec.	Rec.	F1
Random Forest	91%	0.92	0.90	0.91
SVM	86%	0.87	0.85	0.86
KNN	84%	0.85	0.83	0.84
Decision Tree	81%	0.82	0.80	0.81
Naive Bayes	76%	0.77	0.74	0.75

Random Forest achieved the highest accuracy of 91%, attributable to its ensemble architecture. SVM performed competitively at 86%, while KNN recorded 84%. Decision Tree achieved 81% and Naive Bayes recorded the lowest accuracy of 76%.

B. SHAP Feature Importance Analysis

Table II presents the top 10 features ranked by mean absolute SHAP value from the Random Forest model.

TABLE II. TOP PREDICTORS BY MEAN |SHAP| VALUE

Rank	Feature	Category	|SHAP|
1	Assignment Submission Rate	Behavioural	0.48
2	Lab Attendance (%)	Behavioural	0.43
3	Daily Self-Study Hours	Behavioural	0.39
4	Prior Semester GPA	Academic	0.36
5	Lecture Attendance (%)	Behavioural	0.31
6	Motivation Level	Psychological	0.27
7	No. of Backlogs	Academic	0.24
8	Project Participation	Engagement	0.21
9	Stress Index	Psychological	0.18
10	Peer Study Group	Engagement	0.14

The three most influential predictors are all behavioural: assignment submission rate (0.48), laboratory attendance (0.43), and daily self-study hours (0.39). Notably, these surpass prior semester GPA (0.36), challenging the conventional assumption that past grades are the primary determinant of future performance.

VII. DISCUSSION

The results provide compelling empirical support for ML in educational analytics. The SHAP analysis demonstrates that behavioural engagement indicators are stronger predictors of academic outcomes than prior academic performance for engineering students. This finding is particularly meaningful in engineering education, where cumulative practical engagement is central to learning progression.

The superiority of Random Forest (91%) over individual classifiers confirms the advantage of ensemble learning for multi-dimensional educational datasets, consistent with prior research [2][5]. The relative importance of laboratory attendance (0.43) over lecture attendance (0.31) reflects the practical, skills-oriented nature of engineering learning.

From a practical standpoint, these findings support the development of targeted early warning systems in engineering departments that monitor behavioural indicators in real time. Faculty advisors could be alerted when a student's assignment submission rate or laboratory attendance falls below a defined threshold, enabling proactive and personalised intervention.

VIII. CONCLUSION

This study demonstrates that machine learning models, particularly Random Forest, can effectively predict undergraduate engineering student academic performance with 91% accuracy. SHAP-based analysis established that behavioural predictors—assignment submission rate, laboratory attendance, and daily self-study hours—are the strongest determinants, surpassing prior academic performance in predictive significance.

These findings advocate for behavioural monitoring as the cornerstone of early intervention systems in engineering programmes. Future research should focus on real-time deployment, cross-institutional validation, and the development of educator-facing dashboards that translate model outputs into practical alerts and recommendations.

REFERENCES

[1] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," Kybernetes, vol. 51, no. 2, pp. 916–934, 2022.

[2] P. Dabhade et al., "Educational data mining for predicting students' academic performance using machine learning algorithms," Materials Today: Proceedings, vol. 47, pp. 5260–5267, 2021.

[3] F. A. Orji and J. Vassileva, "Machine learning approach for predicting students' academic performance and study strategies based on their motivation," arXiv preprint arXiv:2210.08178, 2022.

[4] P. Nayak, S. Vaheed, S. Gupta, and N. Mohan, "Predicting students' academic performance by mining educational data through machine learning-based classification model," Education and Information Technologies, vol. 28, no. 11, pp. 14611–14637, 2023.

[5] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," Applied Sciences, vol. 10, no. 3, p. 1042, 2020.

[6] Y. S. Balciğlu and M. Artar, “Predicting academic performance of students with machine learning,” Information Development, vol. 41, no. 3, pp. 896–915, 2025.

[7] S. Rajendran, S. Chamundeswari, and A. A. Sinha, “Predicting the academic performance of middle- and high-school students using machine learning algorithms,” Social Sciences & Humanities Open, vol. 6, no. 1, p. 100357, 2022.

[8] M. Yağcı, “Educational data mining: Prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, p. 11, 2022.

[9] A. Nabil, M. Seyam, and A. Abou-Elfetouh, “Predicting students’ academic performance using machine learning techniques: A literature review,” International Journal of Business Intelligence and Data Mining, vol. 20, no. 4, pp. 456–479, 2022.

[10] S. Oreshin et al., “Implementing a machine learning approach to predicting students’ academic outcomes,” in Proc. 2020 1st Int. Conf. Control, Robotics and Intelligent System, pp. 78–83, IEEE, 2020.