© January 2026 | IJIRT | Volume 12 Issue 8 | ISSN: 2349-6002

Predictive Learning Analytics for Student Performance Using Machine Learning: A Study in Coaching Environments

Gnaneswar T. S
SRM Institute of Science and Technology

Abstract— **Predicting student performance is crucial in modern coaching environments, where learners undergo frequent assessments and accelerated preparation cycles for board examinations and competitive entrance tests. Traditional evaluation methods primarily based on periodic tests and teacher judgement are often subjective and unable to identify low-performing students at an early stage. This paper proposes a comprehensive predictive learning analytics framework using machine learning to forecast student performance based on academic, behavioural, and derived metrics.**
A dataset of 180 students across two academic terms with 32 input features was developed. Machine learning models, including Logistic Regression, Decision Tree, K-Nearest Neighbours (KNN), and Random Forest, were trained and evaluated. Experimental results show that the Random Forest model achieved the highest accuracy of 89.3%, outperforming other models. The proposed approach demonstrates the capability of ML-based predictive analytics to support early intervention, personalised learning strategies, and institutional decision-making in coaching environments.

I. INTRODUCTION

Coaching institutions and supplementary learning environments play a significant role in academic performance, especially in countries with competitive examination systems. These environments generate substantial amounts of assessment data weekly tests, monthly exams, accuracy metrics, attendance patterns yet most institutes fail to utilise them for data- driven insights.
Students who start falling behind early often remain unnoticed until late in the academic cycle. Manual analysis becomes impractical when batch sizes exceed 50–100, limiting teachers' ability to track long-term learner progress.
Predictive Learning Analytics (PLA) integrates data science, educational psychology, and machine learning to extract actionable insights from learner data. PLA allows instructors to systematically identify weak learners, track their improvement patterns, and personalize teaching.
This research presents a machine learning framework tailored for coaching environments using realistic student performance indicators.

II. BACKGROUND AND MOTIVATION

A. Traditional Challenges in Coaching Systems

1. Teacher judgement is subjective
2. Large volumes of test data remain unused
3. No early-warning mechanism for weak students
4. Behavioural indicators (attendance, homework) rarely analysed
5. Inconsistent tracking of improvement trends

B. Need for Predictive Learning Analytics

Predictive analytics helps coaching institutes:

- Forecast student outcomes scientifically
- Detect potential dropout or failure risks
- Provide personalised academic guidance
- Reduce the teacher's manual workload
- Improve institutional success rates

III. PROBLEM STATEMENT

Existing academic systems in coaching environments lack:

1. Automated tools to predict student performance early
2. Multi-dimensional analysis of academic + behavioural data
3. Data-driven insights to identify weak learners

4. Model comparison frameworks to determine best ML fit
5. Interpretability for teachers to understand predictions

Thus, a machine learning-based predictive framework is required to address these gaps.

IV. OBJECTIVES OF THE STUDY

The major objectives are:

1. To design a predictive learning analytics model using multiple machine learning algorithms.
2. To construct a realistic dataset representing academic and behavioural indicators.
3. To compare the performance of ML models.
4. To identify critical features that contribute to student performance.
5. To create an early-warning system for weak student identification.
6. To develop a generalisable framework that can be adopted by any coaching institution.

V. LITERATURE REVIEW

C. Traditional Student Performance Prediction

Studies commonly used:

- Regression techniques
- Naïve Bayes
- Basic Decision Trees
- Support Vector Machines

Most work is limited to university datasets (GPA, semester marks, attendance).

D. Machine Learning in Education

Recent models apply Random Forest, Gradient Boosting, and Neural Networks. However:

- Behavioural features remain under-used
- Most models rely solely on exam marks
- Limited focus on coaching institution datasets

E. Research Gaps Identified

1. Lack of studies using multi-dimensional data
2. No clear comparison of classical ML models
3. No framework focusing on supplementary coaching environments
4. Limited early-warning prediction systems

This study fills these gaps with a comprehensive and realistic approach.

VI. METHODOLOGY

F. Dataset Description

A total of 180 students from a coaching environment across two terms were considered. Features: 32 input features grouped into 3 categories:

1) Academic Indicators

- Weekly test scores (8 tests)
- Monthly exam scores (2 tests)
- Final-term exam marks
- Subject-wise mastery
- Question accuracy

2) Behavioural Indicators

- Attendance percentage
- Homework completion rate
- Doubts asked per week
- Submission punctuality
- Class participation score

3) Derived Indicators

- Performance stability index
- Improvement slope
- Error repetition frequency
- Time per question

Output Variable:

Performance Class: Low / Medium / High

G. Data Preprocessing

Steps include:

1. Handling missing values
2. Removal of outliers
3. Normalisation using Min-Max scaling
4. Encoding performance classes
5. Train-test split: 80% training, 20% testing

H. Machine Learning Algorithms Used

1) Logistic Regression

A statistical model predicting probability of performance classes.

2) Decision Tree

Tree-like model dividing students based on criteria (attendance > 80%, accuracy > 70%, etc.).

3) K-Nearest Neighbours

Based on similarity with other students.

4) Random Forest

Ensemble of multiple decision trees. Outperformed all models due to:

- Non-linear handling

- Feature importance accuracy
- Resistance to overfitting

VII. EVALUATION METRICS

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

VIII. RESULTS AND DISCUSSIONS

I. Model Comparison Table

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.5%	0.76	0.74	0.75
Decision Tree	81.1%	0.79	0.78	0.78
K-Nearest Neighbours	74.6%	0.71	0.70	0.70
Random Forest	89.3%	0.87	0.86	0.86

J. Feature Importance (Top Factors)

1. Weekly test average
2. Accuracy rate
3. Attendance percentage
4. Homework completion
5. Improvement slope
6. Error repetition frequency

Interpretation:
High regularity and consistent practice is the strongest indicator of performance

K. Discussion

Random Forest outperformed all other models due to its:

- Ensemble nature
- Stability against noise
- Effective feature sampling
- Ability to combine behavioural + academic features

Logistic Regression and KNN were simpler but performed weaker due to inability to capture non-linear patterns.

IX. CONCLUSION

This study successfully demonstrates that machine learning can accurately predict student performance in coaching environments. The Random Forest model achieved the highest accuracy of 89.3%, showing strong predictive capability. The integration of behavioral features significantly improved prediction accuracy. The framework is highly beneficial for early identification of weak learners and targeted interventions.

X. FUTURE SCOPE

1. Deep learning models (LSTM, GRU)
2. Real-time dashboards for teachers
3. Inclusion of psychological factors
4. Multi-institution dataset generalisation
5. Integration with online learning platforms (LMS)
6. Reinforcement learning for personalised study plans

REFERENCES

[1] H. Altabrawee et al., “Predicting Students’ Performance Using Machine Learning Techniques,” 2019.

[2] Priya S. et al., “Student Performance Prediction Using Machine Learning,” SRM Institute.

[3] A. F. Heikal, “Student Performance Prediction: A Comprehensive Analysis,” ASRIC,2023.

[4] J. Wang & Y. Yu, “Machine Learning in Online Learning Analytics,” PLOS ONE,2025.