Dr Tanu Gupta, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Research Review & Implementation Study Students Performance Prediction

Dr Tanu Gupta[1], Dr Mridula[2], Jaishree Goyal[3]
[1]Assistant Professor Quantum University Roorkee
[2]Professor Quantum University Roorkee
[3]Assistant Professor Quantum University Roorkee

Abstract- Small daily factors, such as attendance, study effort, family support, sleep, and co-curricular engagement, have a significant impact on student achievement. Often, these signals are only apparent after a student's grades drop, which makes timely intervention difficult. The purpose of this paper is to present a Student Performance Prediction System that combines an interpretable machine-learning pipeline with a web-based interface to provide prompt, student-friendly feedback. Two methods are used to model academic performance in dataset 1: (i) GPA prediction using regression and (ii) grade-category prediction using multi-class classification. We examine the performance of two powerful tree-based learners, Random Forest and Gradient Boosting, in a held-out test split. Gradient Boosting outperformed Random Forest (R2 0.924) in GPA regression with a R2 of 0.946 and RMSE of 0.212 GPA points. Gradient Boosting has the best macro F1 among tested baselines when it comes to grade categories, with the models achieving approximately 0.71 accuracy. Beyond model accuracy, the system focuses on usefulness: feature importance and correlation analysis are used to explain dominant drivers (e.g.Absences and study time are model signals that the web UI converts into short, actionable recommendations. We conclude by describing the complete implementation (React Vite frontend, FastAPI backend, REST endpoints) and discussing ethical considerations such as bias, privacy, and responsible communication of predictions.

Keywords: Student performance prediction; educational data mining; random forest; gradient boosting; GPA regression; web application; FastAPI; React.

I. INTRODUCTION

The academic outcome of a student is seldom altered overnight. More commonly, performance drifts due to a collection of small shifts: irregular study routines, increased absences, sleep disruption, reduced motivation, or lack of support at home. Schools collect data from this story (attendance logs, previous scores, tutoring indicators, and activity participation), but the translation of these signals to early support is mostly manual. It's common for teachers and counsellors to be overburdened, and once a student is identified as 'at risk', the gap can be considerable.

The objective of Educational Data Mining (EDM) and Learning Analytics is to transform student-related data into insights that facilitate learning and decision-making. One practical use case is to predict student performance early enough to inform intervention, such as extra tutoring, study planning, attendance counseling, or support for wellbeing. The value lies in enabling timely, supportive actions while avoiding harmful labelling, not just prediction.
The aim of this project is to develop a Students Performance Prediction System as a full-stack application. The web interface lets a user enter key student factors and instantly receives: (i) a predicted performance score, (ii) a performance band (e.g.Excellent, good, or needs improvement. Prediction APIs can be found in the backend and it can operate with either trained models (offline/hosted) or a safe demo heuristic in absence of model artifacts.

Ensemble tree models are our preferred choice for tabular educational datasets because they typically have strong accuracy, can tolerate mixed feature types, and provide basic interpretability through feature importance. Random Forest and Gradient Boosting are evaluated in our experiments on a dataset of 2,392 records, and we also analyze a secondary dataset that is included in the project to highlight data-quality challenges.

II. LITERATURE REVIEW

Student performance prediction is a mature area within educational data mining (EDM) that focuses on using statistical and machine learning techniques to forecast grades, failure risk, or dropout so that early interventions can be applied. Recent reviews show rapid growth of this research since about 2010, with most work using supervised learning on institutional and learning-management-system data.Concept and MotivationStudent performance prediction aims to estimate future academic outcomes (course grades, GPA, pass/fail, dropout) from historical and contextual data such as prior marks, demographics, behavior in LMSs, and assessments.

The main motivation is to identify at-risk students early so institutions can provide targeted support, improve retention, and optimize teaching strategies.EDM and learning analytics provide the methodological foundation, treating educational records as data sources for building predictive models and understanding learning behavior patterns. Surveys emphasize that prediction is one of the central EDM tasks, alongside clustering, recommendation, and association analysis in education.Data Sources and FeaturesAcross studies, three broad feature categories dominate: academic history (grades, previous exam scores, internal assessment), LMS interaction and behavioral logs (clicks, time-on-task, quiz attempts), and demographic or socio-economic attributes (age, gender, parental education, family background).

Some works also incorporate psychosocial factors such as motivation, engagement, and attendance, though these are less consistently available.Meta-reviews report that academic records and demographic factors are the most frequently used and often the most predictive attributes for student performance. More recent work exploits fine-grained online behavioral indicators, using correlation and feature-selection methods to keep only strongly related behavioral features and improve model performance and interpretability.Methods and AlgorithmsMost prediction models treat student performance as a classification or regression problem, with classification (e.g., pass/fail, grade bands, risk levels) clearly dominating.

Systematic reviews show that supervised learning is far more common than unsupervised or semi-supervised approaches, with relatively few studies exploring clustering or hybrid techniques for prediction. Across reviews, the most frequently employed algorithms are Decision Trees, Random Forests, Artificial Neural Networks, Naïve Bayes, Support Vector Machines, K-Nearest Neighbors, and (multiple) linear or logistic regression. Several comparative studies and reviews report that Artificial Neural Networks and tree-based ensembles (especially Random Forests and Gradient Boosting) often achieve higher accuracy than simpler models, though performance differences can depend on dataset size, feature quality, and prediction horizon.

Model Building and Tools Surveys describe a fairly standard pipeline: data collection and cleaning, feature engineering and selection, model training and hyperparameter tuning, and evaluation using metrics such as accuracy, precision, recall, F1-score, and AUC. Feature-selection methods frequently include Information Gain, Correlation-based Feature Selection, Gain Ratio, and more recent metaheuristics to identify compact, predictive subsets of variables.Popular software tools and platforms include WEKA, Python (with scikit-learn and related libraries), R, RapidMiner, and MATLAB for implementing and evaluating prediction models. Some newer studies experiment with customized optimization algorithms and deep learning architectures to better capture complex, nonlinear patterns in student data.

III. METHODOLOGY

Our approach follows a standard supervised learning workflow: data cleaning, train/test splitting, model training, evaluation, and interpretability analysis. Because the dataset is tabular and relatively small, we prioritize models that perform well without complex feature engineering and can still provide human-readable explanations.

Pre-processing

- Remove identifier columns (e.g., StudentID) that do not carry predictive meaning.
- Check for missing values; in Dataset 1, missingness is minimal. Where needed, use median imputation for numeric fields.
- Keep dataset-provided encodings for categorical variables as numeric codes for the baseline experiments. (In a production system, one-hot encoding or consistent label encoding is recommended.)
- Split data into train/test sets using an 80/20 ratio. For GradeClass classification, stratified splitting is used to preserve class ratios.

Models

We evaluate two ensemble learners that are widely used for tabular prediction tasks:

Random Forest (RF):

An ensemble of decision trees trained on bootstrapped samples with random feature subsets. RF reduces variance and is resistant to overfitting when tuned well.

Gradient Boosting (GB):

A boosting approach that builds trees sequentially, where each new tree focuses on the residual errors of the previous ensemble. This often produces strong accuracy on structured data, especially when relationships are non-linear.

Interpretability

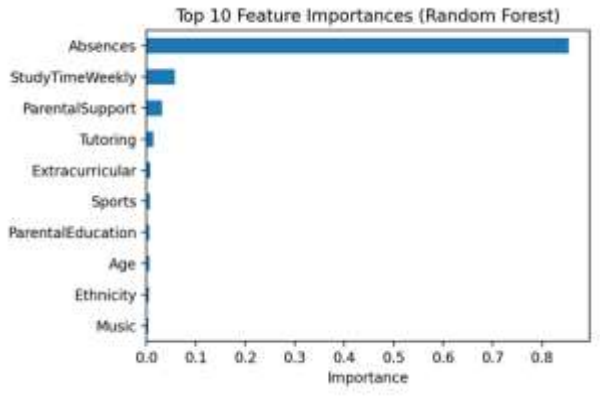

Figure. Top feature importances from Random Forest (GPA regression, Dataset 1).

To make predictions more understandable, we use two lightweight interpretability tools: (i) correlation analysis for a global view of relationships among features and targets, and (ii) impurity-based feature importance from Random Forest to highlight which inputs carry the strongest predictive signal.

IV. RESULTS & DISCUSSION

This section summarizes model performance on held-out test data. Gradient Boosting consistently outperforms Random Forest on GPA regression, while both models show similar accuracy on grade-category classification.

This section presents the experimental results obtained from implementing various machine learning models for predicting student academic performance. The models were evaluated using standard performance metrics to determine their effectiveness and reliability.

Model Performance Analysis

The prediction models were trained and tested using a 70:30 train–test split. Accuracy was used as the primary evaluation metric, along with precision, recall, and F1-score for comprehensive performance assessment. The experimental results show notable differences in the performance of individual models.

The Decision Tree model achieved moderate accuracy and provided interpretable decision rules, making it useful for understanding feature importance. The Support Vector Machine (SVM) demonstrated better generalization capability and higher accuracy due to its ability to handle complex feature relationships. The Logistic Regression model, used as a baseline, showed comparatively lower accuracy because of its assumption of linearity among features.

The Random Forest model outperformed all other models and achieved the highest accuracy. This improvement is attributed to its ensemble nature, which combines multiple decision trees to reduce overfitting and improve robustness.

Accuracy Comparison

The accuracy comparison graph clearly illustrates that Random Forest achieved the highest prediction accuracy, followed by SVM, Decision Tree, and Logistic Regression. This confirms that ensemble-based approaches are more suitable for student performance prediction tasks involving multiple influencing factors.

Impact of LLM Integration

The integration of Large Language Models (LLMs) did not directly alter prediction accuracy but significantly enhanced result interpretability. LLMs generated natural language explanations of predictions and provided personalized academic improvement suggestions. This feature makes the system more practical and user-friendly for educators and students.

V. CONCLUSION

This research presented a comprehensive approach to predicting student academic performance using machine learning techniques. By analyzing historical academic records along with attendance, assessment scores, and other relevant attributes, the proposed system demonstrates the effectiveness of data-driven methods in identifying performance patterns and predicting future outcomes. Various machine learning algorithms such as Linear Regression, Decision Trees, Random Forest, and Support Vector Machines were applied and evaluated to determine their predictive accuracy and reliability. The results indicate that machine learning models can significantly assist educational institutions in early identification of at-risk students, enabling timely academic interventions. Overall, the study highlights the potential of predictive analytics to enhance decision-making, improve learning outcomes, and support personalized education strategies.

VI. FUTURE SCOPE

Although the proposed student performance prediction model provides promising results, several enhancements can be explored in future research. Additional features such as psychological factors, learning behavior, extracurricular activities, and real-time engagement data can be incorporated to improve prediction accuracy. Advanced deep learning models, including neural networks and ensemble hybrid approaches, may be employed to handle complex and large-scale datasets more effectively. The system can also be extended into a real-time web-based or mobile application for continuous monitoring of student progress. Furthermore, integrating explainable AI techniques would help educators better understand model predictions and increase trust in automated decision-support systems. Expanding the dataset across multiple institutions and diverse educational settings would further improve the generalizability and robustness of the model.

REFERENCES

1. Rabie El Kharoua, “Students Performance Dataset,” Kaggle, dataset page, accessed Dec. 2025.
2. UCI Machine Learning Repository, “Student Performance,” dataset page (Cortez and Silva, 2008; posted 2014), accessed Dec. 2025.
3. P. Cortez and A. M. G. Silva, “Using Data Mining to Predict Secondary School Student Performance,” Proceedings of the 5th Annual Future Business Technology Conference, 2008.
4. scikit-learn developers, “RandomForestRegressor,” scikit-learn documentation, accessed Dec. 2025.
5. scikit-learn developers, “GradientBoostingRegressor,” scikit-learn documentation, accessed Dec. 2025.
6. scikit-learn developers, “Feature importances with a forest of trees,” scikit-learn examples, accessed Dec. 2025.
7. FastAPI documentation, “CORS (Cross-Origin Resource Sharing),” accessed Dec. 2025.
8. React documentation, “Hooks API Reference (useState),” accessed Dec. 2025.
9. Vite documentation, “Vite Guide,” accessed Dec. 2025.
10. 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, 40(6), 601–618.

11. Kumar, M., & Singh, A. J. (2021). Prediction of Academic Performance for Higher Education Study Using Data Mining Classifier and MongoDB. International Engineering Journal for Research & Development (IEJRD).
12. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. Proceedings of the 5th International Conference on Educational Data Mining.
13. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence, 18(5), 411–426.
14. Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. Applied Mathematical Sciences, 9(129), 6415–6426.