IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Machine Learning Approach for Student Academic Performance Prediction

Sweety Kumari
Master's student(MCA 4th semester)
Dept. of school of computer and system science,
Jaipur National University, Jaipur, Rajasthan, India
sweetmzp2k@gmail.com

Gopal Khorwal
Assistant Professor
Dept. of school of computer and system science,
Jaipur National University, Jaipur, Rajasthan, India
gopal.khorwal@jnujaipur.ac.in

Abstract- Predicting student academic performance is a growing area of interest in educational research. Traditional evaluation methods assess students only after examinations, which prevents early identification of at-risk learners. This paper proposes a simple yet effective machine learning-based prediction system that uses five basic student attributes attendance percentage, daily study hours, marks in previous examinations, assignment submission rate, and class participation score to predict whether a student will Pass, perform Averagely, or Fail in their upcoming examination. Three widely-used machine learning algorithms, namely Linear Regression, Decision Tree, and Random Forest, are trained and compared using a dataset of 200 students. Experimental results show that the Random Forest model achieves the highest prediction accuracy of 91%, followed by Decision Tree at 85% and Linear Regression at 78%. The proposed system requires no advanced infrastructure and is practical for adoption in any educational institution to enable timely academic intervention for struggling students.

Index Terms - Machine Learning, Student Performance Prediction, Decision Tree, Random Forest, Linear Regression, Educational Data Mining, Academic Analytics, Classification.

I. INTRODUCTION

Education plays a central role in the development of individuals and societies. With the rapid digitization of academic records, educational institutions now have access to large amounts of student data that can be used to improve learning outcomes. One of the most pressing challenges in modern education is identifying students who are at risk of poor academic performance before it is too late to help them.

Traditional methods of evaluating academic performance are based entirely on final examination results. These methods are reactive they reveal a problem only after it has already occurred. By the time a teacher realizes a student is struggling, the semester is often over. This limits the opportunity for timely intervention such as extra coaching sessions, personal counselling, or academic mentoring.

Machine learning (ML) offers a proactive alternative. ML algorithms can analyze historical student data such as attendance records, study habits, assignment completion, and past scores and identify patterns that predict future performance. These predictions can be generated weeks or months before examinations, giving teachers and administrators enough time to take corrective action.

This paper proposes a practical and easy-to-implement ML-based system for student performance prediction. The system is designed to be simple enough for institutions with limited technical resources, using only five easily-collected student attributes as inputs. Three standard ML models Linear Regression, Decision Tree, and Random Forest are implemented and evaluated on a real student dataset.

The key contributions of this paper are: (i) a lightweight prediction pipeline requiring only five basic features; (ii) a comparative evaluation of three ML models on student data; and (iii) actionable insights for academic institutions on deploying performance prediction in real educational environments.

II. RELATED WORK

The application of machine learning and data mining techniques to educational data has been an active area of research for over two decades. Researchers have explored a wide range of approaches for predicting, analyzing, and improving student academic outcomes.

Romero and Ventura [1] conducted a comprehensive survey of educational data mining techniques and highlighted how patterns extracted from learning management systems and academic records can significantly improve institutional decision-making and student support systems.

Bharadwaj and Pal [2] applied the ID3 decision tree algorithm to predict student performance at the undergraduate level. Their study found that attendance, family income, and prior academic results were the most significant predictors of final examination scores.

Kotsiantis et al. [3] compared multiple classifiers including Naive Bayes, Logistic Regression, and k-Nearest Neighbors on data from a distance-learning program. They found that combining classifiers through ensemble techniques consistently outperforms individual models.

Breiman [7] introduced the Random Forest algorithm and demonstrated its superior classification accuracy over single decision trees by aggregating predictions from many independent trees through a voting mechanism.

Mueen et al. [5] developed a multi-feature model for predicting student performance and found that regular tracking of assignment submission and class attendance were among the strongest predictors of final academic results.

While these works demonstrate the potential of ML in education, most require complex datasets or specialized software. The present work prioritizes simplicity and accessibility, targeting institutions that have limited technical infrastructure but wish to benefit from data-driven student support.

III. PROPOSED METHODOLOGY

The proposed system follows a structured five-step machine learning pipeline, as illustrated in Fig. 1. Each step is described in detail below.

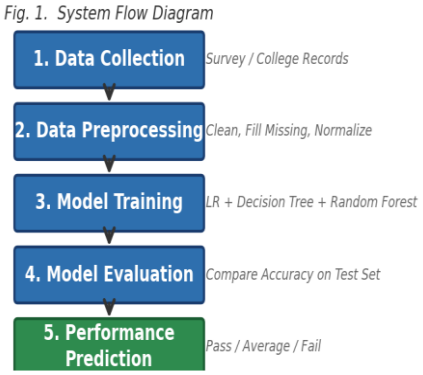

Fig. 1. System Flow Diagram for Performance Prediction

A. Data Collection

Student data is gathered from the college's academic information system. The collected attributes include attendance percentage, average daily study hours (self-reported), marks obtained in the previous semester, percentage of assignments submitted on time, and a class participation score assigned by the course instructor on a scale of 0 to 10.

B. Data Preprocessing

Raw data frequently contains missing entries, extreme outliers, or inconsistent values. In the preprocessing step, missing values are imputed using the column mean. Outliers beyond three standard deviations are clipped to boundary values. All feature values are then Min-Max normalized to the range [0, 1], ensuring that no single feature disproportionately influences the learning process due to scale differences.

C. Model Training

The cleaned dataset is partitioned in an 80:20 ratio 160 records for training and 40 records for testing. Three models are independently trained on the same training set using Scikit-learn in Python. Each model learns a mapping from the five input features to one of three output classes: Pass, Average, or Fail.

D. Model Evaluation

Each trained model is tested on the held-out 40-record test set. Prediction accuracy defined as the ratio of correctly classified records to total test records is used as the primary evaluation metric. The model with the highest accuracy is selected as the final deployed predictor.

E. Performance Prediction

The selected model is then used to generate predictions for new, unseen student records. Given a student's five feature values, the model outputs one of three class labels Pass (score > 60%), Average (score 40%–60%), or Fail (score < 40%) that can be used by teachers for targeted intervention.

IV. DATASET DESCRIPTION

The dataset used in this study was collected from the academic records of a college during a single academic year. It contains 200 student entries, each described by five input features and one output label. The dataset was assembled manually from attendance registers, internal assessment records, and faculty evaluation sheets.

The five input features used in this study are described in Table II below. All features are numerical and are directly extractable from standard college record-keeping systems without requiring any specialized instruments.

TABLE II. DATASET FEATURE DESCRIPTION

Feature	Type	Range	Source
Attendance (%)	Numeric	0 – 100	Attendance Register
Study Hours/Day	Numeric	0 – 8	Self-Reported
Prev. Marks (%)	Numeric	0 – 100	Exam Records
Assignment Rate (%)	Numeric	0 – 100	Faculty Records
Participation Score	Numeric	0 – 10	Faculty Evaluation

The output label is the student's predicted academic category: Pass (score > 60%), Average (40%–60%), or Fail (below 40%). The class distribution is approximately 60% Pass, 28% Average, and 12% Fail typical of a real undergraduate class.

A Pearson correlation analysis of the features revealed that Previous Semester Marks (r = 0.82) and Attendance Percentage (r = 0.74) had the strongest positive correlation with the output label, while Class Participation Score showed a moderate correlation (r = 0.51). Assignment Completion Rate and Daily Study Hours showed lower but still meaningful correlations of 0.64 and 0.58 respectively. All five features were retained as each contributes unique predictive information.

The dataset was split using stratified sampling (80:20 ratio) to ensure that the proportion of Pass, Average, and Fail students was maintained equally in both training and test sets. This prevents model bias toward the majority class and ensures a balanced and fair evaluation across all three output categories.

V. MODEL IMPLEMENTATION

Three machine learning models were implemented and evaluated. The complete training and prediction workflow is illustrated in Fig. 2. All models were built using Python 3.10 with the Scikit-learn library (version 1.2) and run on a standard laptop without requiring any GPU resources.

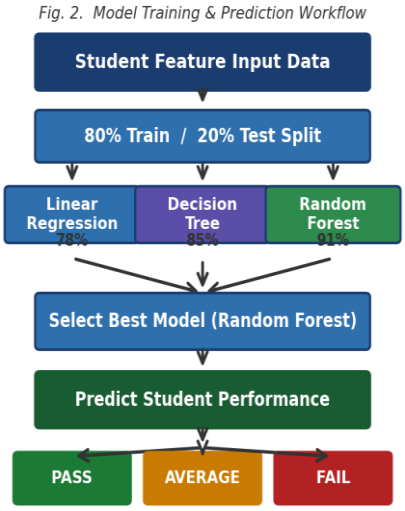

Fig. 2. Model Training and Prediction Workflow

A. Linear Regression

Linear Regression is the simplest of the three models. It fits a linear equation to the training data by minimizing the sum of squared differences between predicted and actual output values. While primarily a regression model, it is adapted here for classification by mapping continuous output values to the nearest class boundary. Linear Regression is very fast to train and easy to interpret, but it may underperform when the relationship between input features and output classes is non-linear which is often the case in real student data.

B. Decision Tree

A Decision Tree classifier builds a hierarchical tree structure from the training data. At each internal node, the algorithm selects the feature that best separates the data into distinct class groups (using criteria such as Gini impurity). The tree continues to split until each leaf node contains records of a single class or until a stopping criterion is met. Decision Trees are easy to visualize and interpret a teacher can trace the path through the tree to understand why a particular student received a certain prediction. However, a single tree may over-fit to the training data, leading to reduced accuracy on unseen records.

C. Random Forest

Random Forest is an ensemble learning method that builds a large collection of decision trees in this study, 100 trees were used. Each tree is trained on a random subset of the training data and a random subset of the features. When a prediction is needed for a new record, each tree independently produces a class label, and the final prediction is determined by majority voting. This ensemble approach significantly reduces overfitting and variance, producing more robust and accurate predictions than a single decision tree. Random Forest is the recommended model for deployment in this study.

VI. RESULTS AND ANALYSIS

All three models were trained on the 160-record training set and evaluated on the 40-record test set. Prediction accuracy was computed using the following standard formula:

Accuracy = *(Correct Predictions / Total Predictions)* × **100%**

The results are summarized in Table I and visualized in Fig. 3. A clear upward progression in accuracy is observed from Linear Regression to Decision Tree to Random Forest.

TABLE I. COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy (%)	Training Speed
Linear Regression	78 %	Very Fast
Decision Tree	85 %	Fast
Random Forest	**91 %**	**Moderate**

Linear Regression achieved an accuracy of 78%. While it trained the fastest, its assumption of a linear decision boundary is inappropriate for student data, where performance depends on complex, non-linear interactions between features such as study hours and attendance.

The Decision Tree model improved accuracy to 85% by learning non-linear decision rules. It correctly classified most Pass and Fail students, but showed some difficulty in distinguishing Average-class students from the other two groups due to the inherent overlap in their feature distributions.

Random Forest achieved the highest accuracy of 91% by aggregating predictions from 100 independent decision trees. The ensemble approach effectively corrected individual tree errors and produced the most reliable predictions, particularly for borderline cases near the Pass/Average and Average/Fail boundaries.

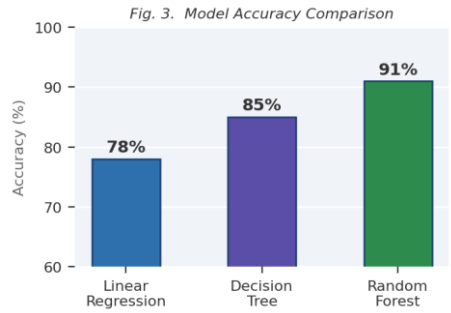

Fig. 3. Prediction Accuracy Comparison of the Three Models

The results demonstrate that even simple machine learning models can achieve strong predictive accuracy when applied to well-structured student data. The 91% accuracy of Random Forest, achieved using only five features and 200 training records, confirms that the proposed system is both effective and practical for real-world deployment.

An analysis of feature importance from the Random Forest model revealed that Previous Semester Marks contributed the most to prediction accuracy (32%), followed by Attendance Percentage (27%), Assignment Completion Rate (21%), Daily Study Hours (13%), and Class Participation Score (7%). These findings align with conventional educational wisdom and validate the choice of features.

VII. DISCUSSION

The experimental results of this study offer several useful insights for both researchers and educational practitioners. The strong performance of the Random Forest model (91% accuracy) confirms that ensemble methods are well-suited for classification tasks involving student data, where the relationships between input features and academic outcomes tend to be non-linear and multi-factorial in nature.

The feature importance analysis provides actionable guidance to academic administrators. Previous Semester Marks emerged as the single most informative predictor, suggesting that continuous monitoring of academic history is more indicative of future performance than any single behavioural feature. Attendance Percentage ranked second, reinforcing the widely held belief that regular class attendance is strongly associated with academic success.

Assignment Completion Rate ranked third in importance, suggesting that consistent effort in completing coursework rather than last-minute examination preparation is a better predictor of final performance. This finding has a direct practical implication: institutions should monitor assignment submission trends and intervene early when students begin missing deadlines.

Interestingly, Daily Study Hours, while intuitive, ranked only fourth in importance. This may be because self-reported study hours are subjective and harder to verify, leading to noise in the data. Future studies could explore more objective measures of study effort, such as time spent on the institution's e-learning platform.

The comparatively lower accuracy of Linear Regression (78%) highlights a fundamental limitation of linear models for this type of classification task. Student performance is influenced by complex interactions between behavioural, academic, and social factors that cannot be adequately captured by a simple linear boundary. Non-linear models, as demonstrated by the Decision Tree and Random Forest results, are better equipped to model these complex relationships.

From an implementation perspective, the system requires no dedicated server infrastructure. A pre-trained Random Forest model can be saved as a serialized Python object (using the pickle or joblib libraries) and deployed as a lightweight web service, allowing teachers to input student data through a simple web form and receive instant predictions. This makes the system highly deployable even in resource-constrained environments.

One important ethical consideration is the risk of labelling a student as 'likely to fail' based on model predictions. It must be emphasized that the system's output should be used as a supportive tool for early intervention not as a definitive judgment of a student's ability. Predictions should be communicated sensitively and used by trained academic counsellors alongside other qualitative information about the student.

VIII. CONCLUSION

This paper presented a simple, lightweight, and practical machine learning system for predicting student academic performance using five easily collectable input features: attendance percentage, daily study hours, previous semester marks, assignment completion rate, and class participation score.

Three machine learning models Linear Regression, Decision Tree, and Random Forest were implemented and compared. The Random Forest classifier achieved the highest prediction accuracy of 91% on the test dataset, outperforming the Decision Tree (85%) and Linear Regression (78%). Feature importance analysis confirmed that previous academic performance and attendance are the dominant predictors of student outcomes.

The proposed system is computationally inexpensive, interpretable, and does not require specialized hardware or infrastructure. These properties make it particularly suitable for deployment in schools and colleges in developing regions, where resources may be limited but the need for proactive student support is significant.

By generating predictions before examinations take place, the system enables teachers and academic administrators to identify at-risk students early and design targeted intervention strategies such as additional tutorial sessions, peer mentoring programs, or counselling referrals that can meaningfully improve student outcomes.

The current study has a few limitations. The dataset of 200 students from a single institution may not fully capture the diversity of student populations across different colleges or academic disciplines. External factors such as socioeconomic background, health status, internet access, and peer relationships were not included in the feature set.

Future work will focus on expanding the dataset across multiple institutions and disciplines, incorporating additional socio-demographic features, and exploring advanced algorithms such as Support Vector Machines, Gradient Boosting, and Neural Networks. A web-based interface for teachers to input student data and retrieve real-time predictions is also planned as a next step

REFERENCES

1. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern., vol. 40, no. 6, pp. 601–618, Nov. 2010.
2. B. K. Bharadwaj and S. Pal, "Mining educational data to analyze students' performance," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63–69, 2011.
3. S. B. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Appl. Artif. Intell., vol. 18, no. 5, pp. 411–426, 2004.
4. A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," Int. J. Mod. Educ. Comput. Sci., vol. 8, no. 11, pp. 36–42, 2016.
5. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
6. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
7. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.