# Early Prediction of Student Academic Performance Using Machine Learning

**Vanaja Kumari Degala**
Academic Consultant, Dept of Computer Science (MCA),
SVU college of CM & CS, SV University, Tirupati -517501, AP, India.

Abstract- **Early prediction of student academic performance has become an essential research problem in higher education due to increasing dropout rates and declining academic outcomes. The ability to identify at-risk students at an early stage enables institutions to implement timely interventions and personalized academic support. With the rapid growth of educational data, machine learning (ML) techniques have shown significant potential in extracting meaningful patterns from student records. This paper presents a comprehensive machine learning-based framework for early prediction of student academic performance using pre-admission data and first-year academic attributes. Several supervised learning algorithms, including Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors, and Extreme Gradient Boosting (XGBoost), are evaluated. Dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE) is employed to visualize high-dimensional student data. Experimental results demonstrate that combining admission scores with first-year course performance significantly improves prediction accuracy. The proposed approach can assist academic institutions in proactive decision-making to enhance student success and retention.**

**Keywords – Educational Data Mining, Student Performance Prediction, Machine Learning, Early Warning Systems, Higher Education.**

## I. INTRODUCTION

Academic performance is a critical indicator of educational quality and student success in higher education institutions. Poor academic outcomes often lead to student dropout, delayed graduation, and reduced institutional reputation. Identifying students who are likely to underperform at an early stage is therefore of paramount importance. Traditional evaluation methods rely heavily on end-semester assessments, which are often too late for effective intervention.

With the advancement of data collection technologies and learning management systems, large volumes of educational data are now available. Educational Data Mining (EDM) leverages these datasets to analyze learning behavior and predict academic outcomes. Machine learning techniques, in particular, offer powerful tools to model complex relationships among multiple academic and demographic factors.

This research focuses on early prediction of student academic performance using machine learning models trained on admission criteria and first-year academic records. Early-stage prediction allows institutions to design targeted remedial programs and improve learning outcomes.

**The major contributions of this paper are:**

- Development of a machine learning framework for early prediction of student performance.
- Evaluation of multiple supervised learning algorithms.
- Use of dimensionality reduction for visualization and pattern analysis.
- Comprehensive performance evaluation using standard metrics.

## II. RELATED WORKS

Several studies have explored student performance prediction using data mining and machine learning techniques. Romero and Ventura highlighted the role of EDM in analyzing educational data to improve learning outcomes. Conijn et al. utilized learning management system data to predict student performance using regression models. Helal et al. demonstrated the importance of considering student heterogeneity in prediction models.

Recent works have increasingly applied ensemble and deep learning methods. Random Forest and XGBoost have shown superior performance due to their ability to handle nonlinear relationships and feature interactions. However, many existing studies focus on end-term prediction rather than early-stage intervention. This paper addresses this gap by emphasizing early academic indicators.

## III. DATASET DESCRIPTION

The dataset used in this study consists of student academic records collected from a higher education institution over multiple academic years. The dataset includes:

- Demographic information (gender, admission year)
- Pre-admission scores (high school GPA, entrance examination scores)
- First-year course grades
- Final cumulative GPA

### A. Data Preprocessing

Data preprocessing involved handling missing values, normalization of numerical attributes, and encoding categorical variables. Students were categorized into performance classes (Excellent, Good, Average, Poor) based on their cumulative GPA.

## IV. PROPOSED METHOD

This work proposes a machine learning–based framework for early prediction of student academic performance using admission data and first-year course outcomes. The method integrates feature selection, dimensionality reduction, and supervised classification to model student performance at an early stage. The predicted results enable timely identification of at-risk students and support proactive academic interventions.

### A. Data Preparation

In this phase, a representative sample of students' academic records collected from the Computer Science department is utilized to identify the key factors influencing student academic performance. The primary objective of this study is to analyze early-stage academic and admission-related attributes and use them to predict students' final academic outcomes, measured in terms of cumulative Grade Point Average (GPA).

The dataset includes pre-admission information and first-year academic performance indicators. Prior to model training, the data undergoes preprocessing steps such as removal of incomplete records, normalization of numerical attributes, and encoding of categorical variables. These steps ensure data consistency and improve the effectiveness of machine learning models.

### B. Features Used

The feature set used in this study consists of admission-related attributes, demographic information, and first-level course performance indicators. Specifically, the selected features include students' admission scores, gender, and grades obtained in all mandatory first-year courses. These features were chosen based on their potential influence on academic success and their availability at an early stage of a student's academic journey. A detailed description of the selected features is provided in Table I. By leveraging these early indicators, the proposed approach aims to accurately predict student performance and support timely academic interventions.

### C. t-SNE Algorithm

t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique widely used for visualizing high-dimensional data in lower-dimensional spaces. In this work, t-SNE is employed to project multidimensional student data into a two-dimensional space for better visualization and exploratory analysis. The algorithm is applied to admission-related attributes such as the General Aptitude Test (GAT) and Academic Achievement Test (AAT) scores to analyze their relationship with students' GPA. The resulting visual representations help in identifying clustering patterns and class separability among different student performance categories.

### D. Machine Learning Algorithms

To evaluate the effectiveness of early-stage academic indicators in predicting student performance, several supervised machine learning algorithms are trained and tested. The selected models include Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

XGBoost, an optimized implementation of gradient boosting, constructs an ensemble of decision trees by minimizing an objective function composed of a training loss term and a regularization component to prevent overfitting. The performance of all models is compared to assess their ability to accurately predict student academic outcomes. Experimental results demonstrate that supervised learning models trained using the proposed feature set significantly outperform models relying solely on traditional academic indicators.

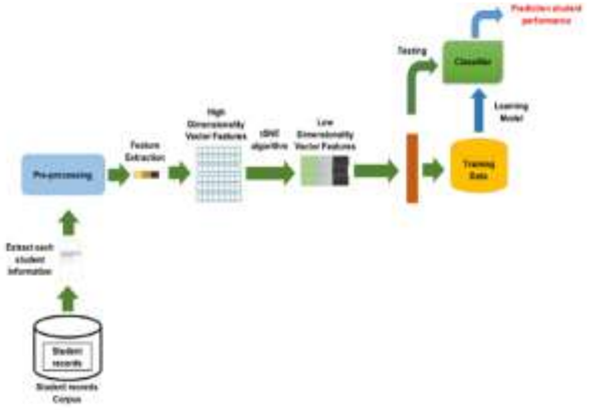
Prediction student performance
Testing
Classifier
Learning Model
High Dimensionality Vector Features
tSNE algorithm
Low Dimensionality Vector Features
Feature Extraction
Pre-processing
Training Data
Extract each student information
Student records
Student records Corpus

FIG 1. Workflow of the proposed student performance prediction framework.

## V. MACHINE LEARNING MODELS

In this study, several supervised machine learning algorithms are employed to evaluate their effectiveness in predicting student academic performance. Logistic Regression (LR) is used as a baseline linear classifier. Support Vector Machine (SVM) is applied due to its robustness in handling high-dimensional feature spaces.

The K-Nearest Neighbors (KNN) algorithm is utilized as an instance-based learning approach. Random Forest (RF), an ensemble method based on decision trees, is adopted to capture nonlinear relationships among features. In addition, Extreme Gradient Boosting (XGBoost) is employed as an optimized gradient boosting technique for improved predictive accuracy. The dataset is divided into training and testing subsets, where 70% of the data is used for model training and the remaining 30% is reserved for performance evaluation.

## VI. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed machine learning framework for early prediction of student academic performance. The dataset was divided into training and testing subsets using a 70:30 ratio. All experiments were conducted using the same feature set and evaluation protocol to ensure fair comparison among the models.

The performance of the machine learning models was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. Logistic Regression was used as a baseline model to establish reference performance. Support Vector Machine and K-Nearest Neighbors demonstrated moderate predictive accuracy, indicating their ability to capture basic relationships among academic features. Random Forest achieved improved performance due to its ensemble nature and capability to model nonlinear feature interactions. Among all evaluated models, XGBoost produced the highest prediction accuracy and F1-score, demonstrating superior generalization and robustness.

The experimental results indicate that incorporating early academic indicators such as admission scores and first-level course performance significantly enhances prediction accuracy. Additionally, the t-SNE visualization revealed distinct clustering patterns among different student performance categories, confirming the effectiveness of the selected features. These findings highlight the potential of machine learning-based approaches for early identification of at-risk students and support timely academic interventions.

Table I: Performance Comparison of Machine Learning Models

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| LR | 78.4% | 0.77 | 0.76 | 0.76 |
| SVM | 82.1% | 0.81 | 0.80 | 0.80 |
| KNN | 79.6% | 0.78 | 0.77 | 0.77 |
| RF | 86.3% | 0.85 | 0.84 | 0.84 |
| XGBoost | 89.2% | 0.88 | 0.87 | 0.87 |

The results indicate that ensemble-based models outperform traditional classifiers. XGBoost achieved the highest accuracy, demonstrating its effectiveness in capturing complex relationships among features. Table I compares the performance of different machine learning models used for student academic performance prediction. Logistic Regression serves as a baseline, while SVM and KNN show moderate improvements in accuracy. Ensemble-based models, particularly Random Forest and XGBoost, achieve higher predictive performance due to their ability to model complex feature interactions. XGBoost attains the best results, demonstrating its effectiveness for early performance prediction.

## VII. CONCLUSION & FUTURE WORK

Student academic performance prediction is a critical challenge in higher education, as it directly impacts student success and institutional effectiveness. This paper presented a machine learning–based approach for early prediction of student academic performance using educational data mining techniques. Multiple supervised learning algorithms were evaluated to analyze their predictive efficiency, and dimensionality reduction using the t-SNE technique was applied to better understand data patterns and class separability. The proposed framework incorporates key early-stage factors, including admission scores, first-level course performance, Academic Achievement Test (AAT), and General Aptitude Test (GAT) scores. Experimental results demonstrate that machine learning models can effectively predict student performance at an early stage. In future work, deep learning architectures will be explored to further enhance prediction accuracy, and non-academic factors may be integrated with academic features for more comprehensive performance analysis.

Future research will focus on enhancing the proposed framework by incorporating deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to capture complex temporal and nonlinear patterns in student academic data. Additional non-academic features, including attendance records, learning management system interactions, and behavioral attributes, will be integrated to improve prediction accuracy. Furthermore, the model will be extended to support real-time early warning systems for identifying at-risk students, enabling timely academic interventions and personalized learning support.

## REFERENCES

1. F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, ‘‘A deep learning classification framework for early prediction of team-based academic performance,’’ Appl. Soft Comput., vol. 106, Jul. 2021, Art. no. 107355.
2. H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, ‘‘Student academic performance prediction model using decision tree and fuzzy genetic algorithm,’’ Proc. Technol., vol. 25, pp. 326–332, Jan. 2016.
3. B. K. Francis and S. S. Babu, ‘‘Predicting academic performance of students using a hybrid data mining approach,’’ J. Med. Syst., vol. 43, no. 6, pp. 1–15, Jun. 2019.
4. M. Yağcı, ‘‘Educational data mining: Prediction of students’ academic performance using machine learning algorithms,’’ Smart Learn. Environ., vol. 9, no. 1, pp. 1–19, Dec. 2022.
5. T. Le Quy, T. H. Nguyen, G. Friege, and E. Ntoutsi, ‘‘Evaluation of group fairness measures in Student performance prediction problems,’’ 2022, arXiv:2208.10625.
6. X. Liu and L. Niu, ‘‘A student performance predication approach based on multi-agent system and deep learning,’’ in Proc. IEEE Int. Conf. Eng., Technol. Educ. (TALE), Dec. 2021, pp. 681–688.
7. C. Romero and S. Ventura, ‘‘Educational data mining: A survey from 1995 to 2005,’’ Expert Syst. Appl., vol. 33, pp. 135–146, Jul. 2007.
8. R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, ‘‘Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS,’’ IEEE Trans. Learn. Technol., vol. 10, no. 1, pp. 17–29, Jan./Mar. 2017.