



STUDENT PERFORMANCE PREDICTION

A Phase 1 Report on Developing a Predictive Model for Bangalore Real Estate

¹MS. KRISHAL RODRIGUES, ²Dr. Shashidhar Kini K

¹Student, ²Professor & Head

¹Department of Master of Computer Applications,
¹Srinivas Institute of Technology, Valchil Mangaluru, Karnataka, India

Abstract : This study focuses on the development of a Student Performance Prediction system using machine learning regression and classification models. The goal is to forecast academic outcomes based on key factors influencing student performance. These factors include demographic attributes, academic history, parental education, socio-economic background, and behavioral indicators. For this purpose, structured data was obtained from public educational datasets and institutional records. The analytical workflow encompasses data acquisition, preprocessing, feature engineering, and the training and evaluation of predictive models such as Logistic Regression and Random Forest. This approach aims to support early intervention strategies in educational settings.

IndexTerms - Student Performance Prediction, Machine Learning, Educational Data Mining, Logistic Regression, Random Forest, Data Preprocessing, Academic Analytics, Predictive Modeling

I. INTRODUCTION

Academic success is a critical determinant of future opportunities and socio-economic mobility for students. In an increasingly competitive and digitized educational landscape, the ability to forecast student performance has gained significant relevance for educators, institutions, policymakers, and parents. Accurate predictions can support early interventions, tailored teaching strategies, and informed decision-making processes that collectively enhance educational outcomes.

While traditional methods of assessing student performance rely heavily on periodic examinations and subjective evaluations, these approaches may not fully capture the diverse factors that influence academic achievement. Numerous intrinsic and extrinsic factors—such as socio-economic status, parental education, attendance, study habits, and psychological well-being—contribute to a student's academic trajectory. However, manually analyzing such multifaceted data can be time-consuming and inconsistent.

With the increasing availability of educational data, machine learning offers a powerful and scalable approach to uncover patterns and generate predictions from historical and real-time student information. This project aims to develop a predictive system that leverages machine learning models to anticipate student academic performance. By doing so, the system seeks to provide actionable insights that can help educators and institutions implement targeted academic support and resource allocation, ultimately fostering a more effective learning environment.

Accurate student performance predictions offer substantial benefits:

- **Students:** Can receive timely academic support and personalized learning plans to improve outcomes
- **Teachers:** Can identify at-risk students early and adapt teaching strategies accordingly
- **Parents:** Can stay informed about their child's academic progress and provide necessary support at home
- **School Administrators:** Can optimize resource allocation, improve curriculum design, and enhance overall institutional performance
- **Policymakers & Educational Planners:** Can use data-driven insights to shape educational policies and intervention programs

The scope of this project is specifically focused on predicting the academic performance of students in secondary and higher secondary levels, using data collected from publicly available educational datasets and institutional academic records [4, 5, 6]. This targeted approach allows the model to reflect the nuanced socio-academic factors influencing student achievement within a localized or institutional context. This paper details the methodology employed in Phase 1 of this project, encompassing data acquisition, preprocessing, analysis, and the proposed machine learning model development strategy, along with anticipated outcomes.

II. EASE OF USE

The proposed student performance prediction system emphasizes simplicity and accessibility for end-users such as educators, academic counselors, and school administrators. Once the machine learning (ML) models are trained, they can be integrated into a user-friendly dashboard or web interface where users input relevant student attributes—such as attendance, prior grades, participation, socio-economic background, and study hours—to obtain immediate performance predictions.

This model eliminates the need for users to possess deep technical knowledge, enabling data-driven decision-making in educational environments with minimal training. The system supports dynamic visualizations and risk indicators to highlight students who may require academic intervention, enhancing early support initiatives.

The architecture is designed for smooth integration with existing Learning Management Systems (LMS) or academic portals via a RESTful API. Real-time insights into student outcomes can be accessed directly through institutional platforms, making the solution practical for large-scale educational deployment.

To ensure optimal usability, the system is developed with both performance and scalability in mind. Lightweight classifiers such as Decision Trees offer real-time inference, while ensemble models like Random Forest or Gradient Boosting enhance predictive accuracy for deeper analysis. The platform is flexible, allowing continuous updates and retraining with new data, ensuring relevance in evolving educational contexts.

Prepare Your Paper Before Styling

Before structuring the final paper in IEEE format, substantial effort was directed toward ensuring comprehensive and coherent content. The initial phase involved collecting raw academic datasets from trusted sources such as Kaggle and UCI Machine Learning Repository. These datasets included diverse student attributes like demographic information, academic scores, behavioral metrics, and parental background.

The preprocessing stage involved cleaning the data by handling missing values, encoding categorical variables such as gender and parental education, and normalizing numerical features like test scores and study time. Feature engineering techniques were employed to derive additional attributes, such as performance trends and attendance ratios, which significantly enhanced model performance.

Various ML models were explored, including Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression was used as a baseline due to its interpretability, while ensemble models were selected for their high accuracy and ability to capture complex patterns. The models were evaluated using metrics such as Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC), ensuring a balanced evaluation of predictive power.

Only after these technical foundations were completed was the report formatted in the IEEE style. The paper was structured logically into sections like Introduction, Literature Review, Methodology, Experimental Results, and Conclusion. Tables, charts, and model evaluation outputs were carefully incorporated, and the document was thoroughly proofread to maintain academic clarity and grammatical accuracy.

2. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. Common abbreviations such as IEEE, SI, and units of measurement (e.g., kg, km, and hrs) do not need to be defined. Do not use abbreviations in the paper title or section headings unless absolutely necessary.

In this paper, the following abbreviations and acronyms are used:

- **ML** – Machine Learning
- **LMS** – Learning Management System
- **API** – Application Programming Interface
- **F1 Score** – Harmonic Mean of Precision and Recall
- **XGBoost** - Extreme Gradient Boosting
- **UCI** – University of California, Irvine
- **SVM** – Support Vector Machine
- **RF** – Random Forest
- **PCA** – Principal Component Analysis
- **CSV** – Comma-Separated Values
- **GPA** - Grade Point Average
- **GUI** – Graphical User Interface
- **KNN** – K-Nearest Neighbors

RESEARCH METHODOLOGY

This section outlines the methodology adopted to develop a machine learning-based predictive model for evaluating student performance. It covers the population and sample, data sources, theoretical framework, and the statistical and machine learning tools employed for analysis and prediction.

3.1 Population and Sample

The population of this study comprises students from primary, secondary, and higher education institutions whose academic performance can be influenced by behavioral, socio-demographic, and academic factors. The focus includes students across various age groups and educational levels represented in publicly available datasets containing academic and socio-economic indicators.

A representative sample of approximately 1,000 to 2,000 student records was selected from well-known datasets, such as the Student Performance Data Set from the UCI Machine Learning Repository (covering Portuguese and mathematics courses) and other public datasets containing academic records, attendance logs, and parental background. These datasets include features such as study time, previous grades, school support, family support, and demographic characteristics.

The sampling strategy ensured diversity in gender, school type, parental education level, and study habits to support supervised classification and regression analysis. Records with excessive missing values or contradictory entries were excluded. Data was balanced using techniques such as SMOTE (Synthetic Minority Oversampling Technique) where necessary to address class imbalance, particularly when classifying performance levels (e.g., pass/fail or grade categories). The data is cross-sectional and reflects recent academic years, ensuring relevance to current educational trends.

3.2 Data and Sources of Data

The study is based on secondary data collected from publicly accessible academic and institutional repositories. The primary sources include the UCI Machine Learning Repository, open Kaggle datasets on student grades, and anonymized academic performance records made available by educational institutions.

Key features extracted from the listings include:

- **Demographic Data** (e.g., age, gender, address type)
- **Academic Background** (e.g., prior grades, class failures, study time)
- **Parental Information** (e.g., education level, job, support at home)
- **Behavioural Indicators** (e.g., absences, health status, free time)
- **Institutional Support** (e.g., school support, tutoring, extracurricular activities)
- **Social and Economic Factors** (e.g., internet access, family relationship quality)

The raw data underwent thorough preprocessing, including handling missing values, normalizing numeric fields (such as study time or absences), and encoding categorical variables (such as school name, gender, and job titles). Feature engineering was performed to derive additional variables like average previous scores or consistency in attendance.

3.3 Theoretical framework

The objective of this research is to construct a model capable of predicting student performance—either as a continuous grade prediction (regression) or as classification (e.g., pass/fail). The dependent variable varies based on the modeling task and includes:

- **Demographic Attributes** (e.g., age, gender)
- **Academic Indicators** (e.g., prior test scores, number of past failures)
- **Study Habits** (e.g., daily study time, extra classes)
- **Parental and Family Context** (e.g., parental education, family support)
- **Behavioural Metrics** (e.g., absences, extracurricular involvement)

Given the non-linear and multi-dimensional nature of these relationships, advanced ML models are employed to capture hidden patterns that traditional models may overlook. Tree-based models and ensemble methods are particularly useful due to their flexibility, interpretability, and performance on structured data.

3.4 Statistical tools and econometric models

This section describes the tools and models employed for analyzing student data and predicting academic performance.

a) Descriptive Statistics

Basic statistical summaries, such as mean, median, mode, and standard deviation, were calculated for all features. These helped assess the data distribution and identify anomalies or outliers, particularly in grades, absences, and study time.

b) Exploratory Data Analysis (EDA)

EDA techniques, including boxplots, correlation heatmaps, and distribution plots, were used to understand variable relationships. For example, study time and parental education level showed positive correlations with performance, while high absences indicated performance risks.

c) Logical Regression

Used as the baseline model for binary classification (e.g., pass vs. fail). Logistic regression provided insights into how each predictor variable influences the likelihood of academic success.

d) Random Forest Classifier

Random Forest was used to model complex feature interactions and non-linear effects. It is robust to overfitting and provides feature importance, aiding interpretability.

e) XGBoost (Extreme Gradient Boosting)

XGBoost was employed due to its high performance and regularization capabilities. It is particularly effective on structured data and was tuned using cross-validation for optimal results.

f) Support Vector Machine (SVM)

SVM was tested to classify students based on hyperplane separation. It is effective for small to medium-sized datasets and handles high-dimensional data well.

g) K-Nearest Neighbors (KNN)

KNN was explored due to its simplicity and effectiveness in classification tasks where class labels depend on similarity in feature space. It performed well in scenarios with clear clusters of performance patterns.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
Attendance	40	100	76	12
Internal Marks	10	50	32	8
External Marks	20	100	65	15
Study Time(hrs.)	0	5	2.5	1.1

Table 4.1 The results from descriptive statistics show that the data is reasonably distributed, with some variability in student performance due to factors such as study time, parental education, and past academic records. Further analysis revealed that previous grades and parental support had a strong influence on final academic performance.

III. ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Project Guide and Head of the Department of MCA, Dr. Shashidhar Kini K, for his invaluable guidance, constant encouragement, and kind support throughout this research work. Appreciation is also extended to the Principal, Dr. Shrinivasa Mayya D, for fostering an environment conducive to completing this project within the institution. The author thanks the management of Srinivas Institute of Technology for their direct and indirect support. Gratitude is also due to all the faculty members and non-teaching staff of the MCA department for their constant help and support. Finally, the author is indebted to parents and friends for their unwavering support and belief throughout this endeavor.

REFERENCES

- [1] Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance.
- [2] Ramesh, V., & Parkavi, D. (2013). Predicting student performance using classification techniques.
- [3] Kotsiantis, S. B. (2012). Use of Machine Learning in Educational Software: A Review.
- [4] Yadav, S. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification.
- [5] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees.

