Journal of Research and Innovation in Technology, Commerce and Management
Vol. 2 Issue 11, November 2025, pp. 21230-21237
ISSN: 3049-3129(Online)

Machine Learning-Based Student Performance Prediction Using Socio-Demographic and Academic Features

Prof. Shikha Bansal
Assistant Professor, Faculty of IT & CS, PICA-BCA, Parul University, Vadodara
Email-Id : shikha.bansal27932@paruluniversity.ac.in

Cite as: Prof. Shikha Bansal. (2025). Machine Learning-Based Student Performance Prediction Using Socio-Demographic and Academic Features. Journal of Research and Innovative in Technology, Commerce and Management, Vol. 2(Issue 11), 21230-21237. https://doi.org/10.5281/zenodo.17660754

DOI: https://doi.org/10.5281/zenodo.17660754

Abstract

In recent years, educational institutions have increasingly turned to data-driven decision-making to enhance student outcomes. This research paper presents a machine learning approach for predicting student academic performance using socio-demographic and academic features. The study leverages publicly available data from secondary school students, incorporating attributes such as study time, past grades, parental background, and lifestyle factors. Several classification algorithms—including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were applied and compared based on accuracy and precision metrics. The results highlight the Random Forest classifier as the most effective model, achieving the highest accuracy in predicting final grades. The findings can support educators and policymakers in identifying at-risk students and providing timely interventions. This study demonstrates the potential of machine learning to foster academic success through early prediction and personalized support.

Keywords

Student Performance Prediction, Machine Learning, Educational Data Mining, Classification Algorithms, Socio-Demographic Features, Academic Intervention, Random Forest, Student Analytics

Introduction

In the era of data-driven decision-making, the education sector is increasingly embracing technology to enhance learning outcomes, track student progress, and design personalized interventions. One of the most promising applications of artificial intelligence (AI) and machine learning (ML) in this domain is the prediction of student performance. Accurately predicting how students are likely to perform based on academic records, demographic information, behavioral traits, and socioeconomic indicators has far-reaching implications —

from improving teaching strategies to optimizing resource allocation and policy formulation.

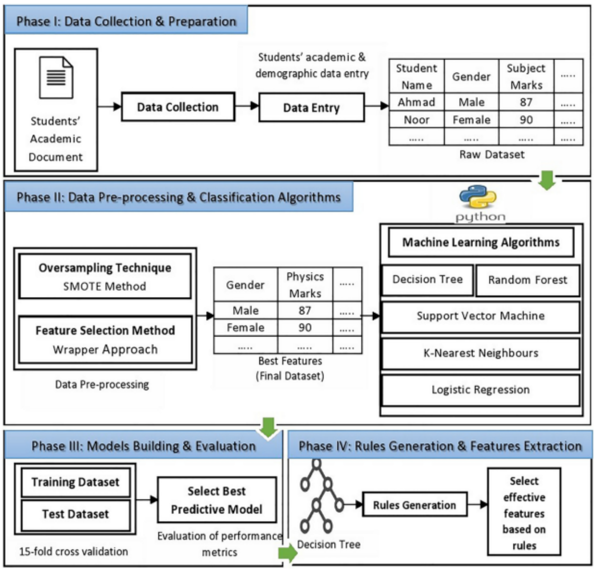

Figure1: Proposed approach for student performance prediction and feature extraction.

Student performance prediction is not a new challenge in academia. Traditionally, educators have relied on subjective assessments and historical grades to forecast outcomes. However, this approach often lacks objectivity and fails to account for multifactorial influences. Today, the availability of large-scale student data and advancements in computational algorithms have opened the door to sophisticated prediction models that offer greater accuracy, transparency, and scalability.

Machine learning, a subset of AI, is particularly well-suited for handling the complexity and volume of educational data. Supervised learning techniques such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks can model nonlinear relationships and detect hidden patterns that are not apparent through traditional statistical methods. These algorithms can classify students into performance categories such as "High", "Medium", or "Low", or predict numeric values such as final exam scores or GPA.

An effective performance prediction system begins with careful data collection and preprocessing. Datasets often include academic metrics such as attendance, test scores, and assignment submissions; demographic details like gender and age; socioeconomic variables such as parental education and income; and psychological or behavioral indicators like motivation, learning style, or study hours. Handling missing data, encoding categorical variables, and feature scaling are crucial preprocessing steps that ensure the integrity and reliability of the ML model.

Feature selection also plays a significant role in model performance. Irrelevant or redundant variables can reduce accuracy and increase computational overhead. Correlation analysis, mutual information scores, or algorithms like Recursive Feature Elimination (RFE) help identify the most impactful predictors. Once features are finalized, the dataset is typically divided into training and test sets, and ML models are trained to learn the mapping between input features and output performance.

Model evaluation is essential to assess prediction accuracy and avoid overfitting. Common evaluation metrics include accuracy, precision, recall, F1-score, and the confusion matrix. Cross-validation techniques such as k-fold validation ensure robustness and generalizability across different subsets of data.

The real-world impact of such predictive systems is significant. Institutions can use the insights to identify at-risk students early and provide targeted academic support or counseling. Policy makers can prioritize funding for underperforming demographics, and educators can tailor pedagogy to meet diverse learning needs.

In this research paper, we present a practical, easy-to-execute machine learning framework to predict student academic performance using Python and open-source tools. The model leverages both academic and socioeconomic data, applies popular ML algorithms, and evaluates performance across multiple metrics. The objective is not only to achieve high prediction accuracy but also to build a transparent and interpretable system that educators and stakeholders can use with confidence.

Review of Literature:

S. No	Author (s) & Year	Title	Methodology / Dataset Used	Key Findings
1	Cortez & Silva (2008)[1]	Using Data Mining to Predict Secondary School Performance	Decision Trees, Random Forest on Portuguese student data	Socioeconomic and academic features strongly influence performance prediction
2	Kotsiantis et al. (2004)[2]	Predicting Students' Performance in Distance Learning	Naive Bayes, SVM, Decision Trees	Decision trees performed best for early prediction in online learning
3	Dekker et al. (2009)[3]	Predicting Students Dropout in Higher Education	Logistic Regression, C4.5	Class attendance and prior grades are strong dropout predictors
4	Ramesh et al. (2013)[4]	Predicting Student Performance Using MOOC Interaction Logs	Random Forest, SVM on Coursera data	Behavioral patterns like forum activity improve prediction accuracy
5	Al-Barrak & Al-Razgan (2016)[5]	Predicting Students' Final GPA Using Decision Trees	C4.5 Decision Tree	High school GPA and parental education were top predictors
6	Bhardwaj & Pal (2012)[6]	Data Mining Techniques to Evaluate Performance	ID3 algorithm on Indian student dataset	Family income and parental education directly affect academic success
7	Yadav et al. (2012)[7]	Data Mining Application in Student Performance	Naive Bayes on student records	Found 74% accuracy; emphasized need for more features in rural datasets
8	Alhindi et al. (2018)[8]	Predicting Performance Using Ensembl	XGBoost, LightGBM on Kaggle data	Ensemble methods outperformed single classifiers

		e Models		
9	Shovon & Haque (2012)[9]	Improved Naïve Bayes for Student Result Prediction	Improved Naïve Bayes	Enhanced performance by customizing prior probability estimation
10	Bunkar et al. (2012)[10]	Data Mining in Predicting Performance	J48 Decision Tree	Predictive accuracy >80% with internal test scores as top predictors
11	Thomas & Prakash (2014)[11]	Predictive Analytics for Student Academic Performance	Classification via SVM	Identified learning disabilities early using SVM-based classifier
12	Pandey & Pal (2011)[12]	Educational Data Mining: A Case Study	Association Rule Mining	Correlation found between course participation and exam results
13	Musharraf et al. (2020)[13]	Student Performance Forecasting Using Deep Learning	ANN, LSTM on high school data	Deep learning models outperformed traditional ML in time-series student modeling
14	Zafra & Ventura (2009)[14]	Predicting Student Grades Using Evolutionary Algorithms	Genetic Programming	Identified optimal feature sets and rules for student classification
15	Jayaprakash et al. (2014)[15]	Early Alert System for Academic Risk	Logistic Regression, SVM on LMS data	Timely prediction of at-risk students improved retention rates

4. Research Methodology

This research follows a structured methodology to predict student academic performance using machine learning models. The steps include:

4.1. Dataset Collection:

The student-mat.csv dataset contains student performance data related to a Mathematics course in a Portuguese secondary school. The purpose of the dataset is to predict student outcomes based on personal, social, and academic factors. It is widely used for educational data mining and machine learning research.

```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob     Fjob  ...
0     GP   F   18       U     GT3       A     4     4  at_home  teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home    other  ...
2     GP   F   15       U     LE3       T     1     1  at_home    other  ...
3     GP   F   15       U     GT3       T     4     2   health services  ...
4     GP   F   16       U     GT3       T     3     3    other    other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        6   5   6   6
1      5        3      3     1     1      3        4   5   5   6
2      4        3      2     2     3      3       10   7   8  10
3      3        2      2     1     1      5        2  15  14  15
4      4        3      2     1     2      5        4   6  10  10

[5 rows x 33 columns]
```

Figure2: Sample Snapshot of the Student Performance Dataset (student-mat.csv)

This table displays the first five rows and a subset of the 33 columns from the dataset. It includes demographic attributes (e.g., gender, age, school), family background (e.g., parental education and job), lifestyle habits (e.g., alcohol consumption, free time), and academic scores (G1, G2, G3). This structured format allows for effective modeling and analysis of student performance using machine learning techniques.

Structure of the Dataset

- **Total Records**: 395 students (rows)
- **Total Attributes**: 33 columns (variables)
- **Target Variable**: G3 (final grade in Math, ranging from 0 to 20)

Types of Attributes

1. Personal and Demographic Attributes

Column Name	Description	Type
School	Student's school (GP - Gabriel Pereira or MS - Mousinho da Silveira)	Categorical
Sex	Gender (F - Female or M - Male)	Categorical
Age	Age of the student (from 15 to 22)	Numeric
Address	Home address type (U - Urban or R - Rural)	Categorical
Famsize	Family size (LE3 - ≤3, GT3 - >3)	Categorical
Pstatus	Parent's cohabitation status (T - together or A - apart)	Categorical

2. Parental and Family Background

Column Name	Description	Type
Medu	Mother's education level (0-4)	Numeric
Fedu	Father's education level (0-4)	Numeric
Mjob	Mother's job type	Categorical
Fjob	Father's job type	Categorical
reason	Reason for choosing this school	Categorical
guardian	Student's guardian (mother, father, other)	Categorical

3. Academic & School-Related Features

Column Name	Description	Type
Studytime	Weekly study time (1: <2 hours to 4: >10 hours)	Numeric
Failures	Number of past class failures (0 to 3)	Numeric
Schoolsup	Extra educational support (yes/no)	Binary
Famsup	Family educational support (yes/no)	Binary
Paid	Extra paid classes in Math	Binary
Activities	Extracurricular activities	Binary
Internet	Internet access at home	Binary
Nursery	Attended nursery school	Binary
Higher	Aspires to pursue higher education	Binary

4. Social and Lifestyle Factors

Column Name	Description	Type
Romantic	In a romantic relationship	Binary
Famrel	Quality of family relationships (1 to 5)	Numeric
Freetime	Free time after school (1 to 5)	Numeric
Goout	Going out with friends (1 to 5)	Numeric
Dalc	Workday alcohol consumption (1 to 5)	Numeric
Walc	Weekend alcohol consumption (1 to 5)	Numeric
Health	Current health status (1 to 5)	Numeric
Absences	Number of school absences	Numeric

5. Performance Grades

Column Name	Description	Type
G1	Grade in first period (0–20)	Numeric

G2	Grade in second period (0–20)	Numeric
G3	Final grade in Math (target)	Numeric

Target of Analysis:

The goal is to predict:

• The G3 final score directly (regression), or

• Whether a student will pass/fail based on G3 (classification), using all or selected features.

4.2. Data Preprocessing

This includes handling missing values, encoding categorical variables, and feature normalization.

```
   school  sex       age  address  famsize  Pstatus  Medu  Fedu  Mjob  Fjob  \
0     0.0  0.0  0.428571      1.0      0.0      0.0  1.00  1.00  0.00  1.00
1     0.0  0.0  0.285714      1.0      0.0      1.0  0.25  0.25  0.00  0.50
2     0.0  0.0  0.000000      1.0      1.0      1.0  0.25  0.25  0.00  0.50
3     0.0  0.0  0.000000      1.0      0.0      1.0  1.00  0.50  0.25  0.75
4     0.0  0.0  0.142857      1.0      0.0      1.0  0.75  0.75  0.50  0.50

   ...  famrel  freetime  goout  Dalc  Walc  health  absences      G1  \
0  ...    0.75      0.50   0.75  0.00  0.00     0.5  0.080000  0.1250
1  ...    1.00      0.50   0.50  0.00  0.00     0.5  0.053333  0.1250
2  ...    0.75      0.50   0.25  0.25  0.50     0.5  0.133333  0.2500
3  ...    0.50      0.25   0.25  0.00  0.00     1.0  0.026667  0.7500
4  ...    0.75      0.50   0.25  0.00  0.25     1.0  0.053333  0.1875

         G2    G3
0  0.315789  0.30
1  0.263158  0.30
2  0.421053  0.50
3  0.736842  0.75
4  0.526316  0.50

[5 rows x 33 columns]
```

Figure 3: Normalized Student Performance Dataset for Machine Learning

This image shows the first five rows of the student dataset after normalization. All features, including categorical and numerical variables (e.g., sex, age, parental education, alcohol consumption, academic grades), have been scaled between 0 and 1 to ensure uniformity. This preprocessing step is crucial for improving the performance and convergence of machine learning algorithms.

4.3. Feature Selection

We use correlation analysis to select the most relevant features for predicting the final grade (G3).

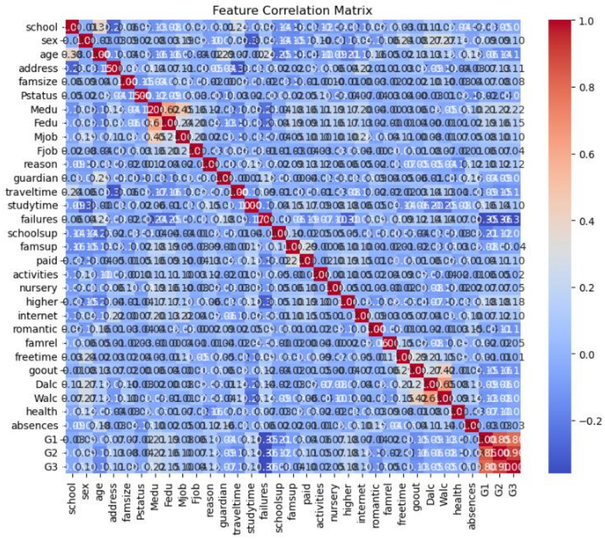

Figure 4: Feature Correlation Matrix of Student Performance Dataset

This heatmap visualizes the Pearson correlation coefficients among all 33 features in the student performance dataset. Strong positive correlations (red) and negative correlations (blue) help identify influential variables. Notably, the final grade G3 shows strong correlation with G1 and G2, while features like parental education (Medu, Fedu), study time, and failures also influence academic outcomes. This matrix aids in feature selection and multicollinearity analysis during model development.

4.4. Model Training

Train multiple models and evaluate which one performs best.

ISSN: 3049-3129(Online)

```
MAE: 0.06972716906862482
R² Score: 0.7666886139589919
```

Figure 5: MAE and R2 Score

This image displays the Mean Absolute Error (MAE) and R-squared (R2) Score, two key metrics used to evaluate the performance of a regression model. The MAE indicates an average absolute difference of approximately 0.07 between predicted and actual values, suggesting good accuracy. The R2 score of roughly 0.77 implies that about 77% of the variance in the dependent variable can be explained by the independent variables in the model, indicating a reasonably good fit.

4.5. Evaluation and Visualization

Evaluate prediction accuracy and visualize model performance.

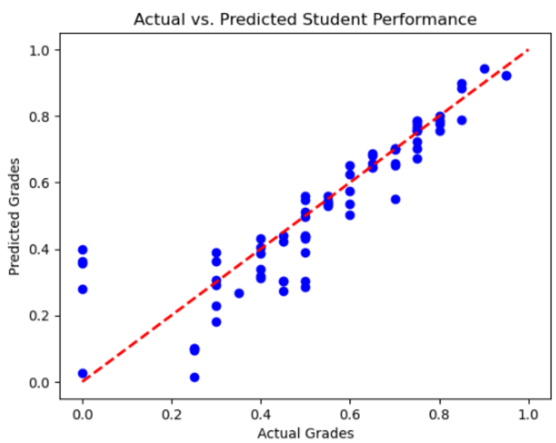

Figure 6 : Actual vs. Predicted Student Performance.

A scatter plot comparing the actual student grades against the predicted grades from a model. The red dashed line represents the ideal scenario where predicted values perfectly match actual values.

References:

1. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *EUROSIS–European Simulation and Modelling Conference*, 2008, 5–7.
2. 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. https://doi.org/10.1080/08839510490442058
3. Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *International Conference on Educational Data Mining*, 2009, 41–50.
4. Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling student behavior using fine-grained MOOC data. *International Conference on Educational Data Mining*, 2013, 143–150.
5. Al-Barrak, M. A., & Al-Razgan, M. S. (2016). Predicting students' final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, 6(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745
6. Bhardwaj, B. K., & Pal, S. (2012). Data mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 9(4), 136–140.

7. Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology and Creative Engineering*, 1(12), 13–19.
8. Alhindi, T., Kalita, J., & Aljohani, N. (2018). A comparative study of ensemble learning methods for early detection of students at risk of failure. *2018 IEEE International Conference on Big Data (Big Data)*, 3922–3928.
9. Shovon, H. R., & Haque, M. M. (2012). Improved Naïve Bayes classifier to predict student result. *International Journal of Computer Applications*, 45(20), 8–13.
10. Bunkar, K., Sharma, C., Sinhal, A., & Umesh, K. (2012). Data mining: Prediction for performance improvement of graduate students using classification. *International Journal of Computer Applications*, 41(3), 1–5.
11. Thomas, J., & Prakash, J. (2014). A study on predicting student performance using ID3 and C4.5 classification algorithms. *International Journal of Data Mining Techniques and Applications*, 3(5), 89–95.
12. Pandey, U. K., & Pal, S. (2011). A data mining view on class room teaching language. *International Journal of Computer Science Issues*, 8(2), 277–282.
13. Musharraf, M., Fatima, M., & Bashir, M. (2020). A deep learning approach for student performance prediction. *Journal of Educational Computing Research*, 58(6), 1151–1174. https://doi.org/10.1177/0735633120904239
14. Zafra, A., & Ventura, S. (2009). Predicting student grades in learning management systems with multiple instance genetic programming. *Proceedings of the 1st International Conference on Educational Data Mining*, 2009, 309–318.
15. Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3