International Institute of Information Technology, Hyderabad

Student Performance Prediction

By- Anjali Suhas Tichkule

Introduction:

Academic performance prediction is crucial in identifying at-risk students early. Machine learning enables data-driven predictions based on student behavior and history. This project leverages ML models to predict final grades and pass/fail status.

Objectives

Predict student's final grade (G3) and classify pass/fail status.

Identify key factors influencing student performance.

Compare different ML models to select the most accurate one.

Provide actionable insights to educators for early intervention.

Early Identification

Predicting performance helps spot at-risk students early.

Challenges

Handling missing or imbalanced data.

Choosing relevant features from socio-academic variables.

Avoiding overfitting while tuning complex models.

Interpreting results for real-world educational use.

Literature Review:

Sr. No.	Study	Year	Description	Limitation
1	Cortez & Silva	2008	Used decision trees and neural nets on Portuguese student data.	Small or localizes dataset
2	Kotsiantis et al.	2004	Found ensemble models outperform single classifiers in student prediction.	Limited feature diversity
3	Bhardwaj & Pal	2011	Emphasized past performance and attendance as key indicators.	Static analysis
4	Yadav et al.	2012	Applied naïve Bayes and decision tree algorithms to educational data.	Lack of real-world deployment
5	Al-Barrak & Al-Razgan	2016	Compared SVM and Random Forest; RF had better accuracy.	Model interpretability

METHODOLOGY

Algorithms

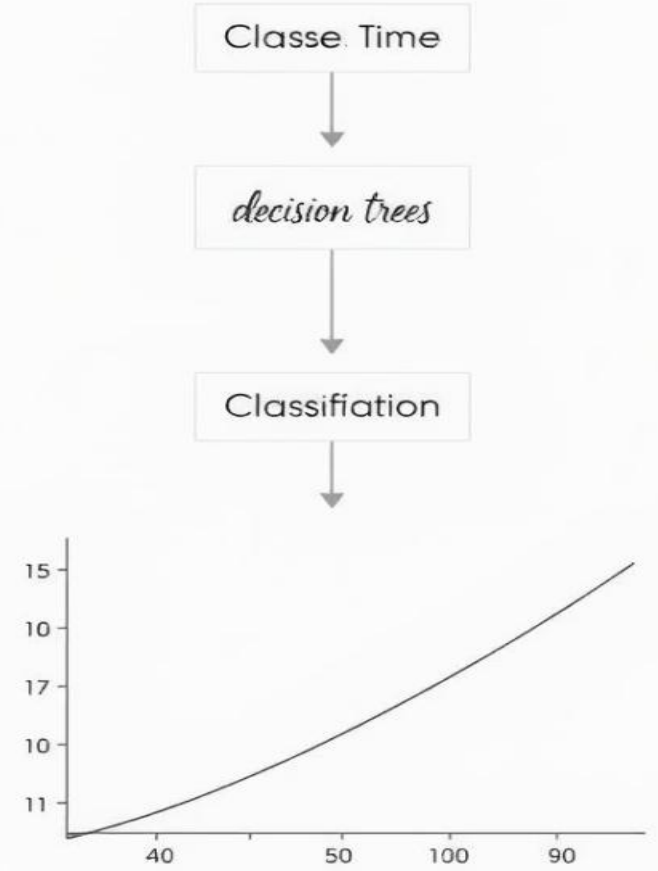
Regression, Classification, Clustering

Model Selection

Logistic Regression, SVM, Random Forest

Evaluation Metrics

Accuracy, Precision, Recall, F1-score with K-fold validation



1. Preprocessed student data by encoding categories and creating a pass/fail label.
2. Trained ML models (Random Forest, XGBoost) and evaluated them to predict grades and classify performance.

Implication

Educational Impact

Supports early intervention strategies for better success.

Accuracy & Limitations

Results impacted by data bias and model generalizability issues.

Future Directions

Expand features and enable real-time prediction capabilities.

1. Limited Generalizability: Using small, localized datasets means the models may not perform well in other educational systems or cultural contexts. This restricts their global or national adoption.
2. Incomplete Student Profiles: Excluding emotional, psychological, or behavioral factors results in models that fail to capture the full picture of student learning, potentially misidentifying at-risk students.
3. Short-Term Focus: Static analysis prevents early detection of performance trends or long-term academic risk, limiting proactive intervention strategies.
4. Lack of Practical Use: Without real-world testing and deployment, research models remain academic exercises and do not translate into tools that help teachers or schools in practice.

Data Collection

The dataset used for this project is the Student Performance Data Set obtained from the UCI Machine Learning Repository.

Source
UCI Machine Learning Repository
Student Performance Dataset

Description
The dataset contains academic performance data of Portuguese secondary school students in two subjects: Mathematics (student-mat.csv) and Portuguese Language (student-por.csv).
For this project, the Mathematics dataset (student-mat.csv) is used.

Data Format: CSV (Comma-Separated Values) with semicolon (;) as delimiter
Total Records: 395
Total Features: 33 attributes per student

Attributes Collected

1. Demographic Informationsex, age, address, famsize, Pstatus
2. Parental and Social BackgroundMedu, Fedu, Mjob, Fjob, reason, guardian.
3. Academic Factorsschoolsup, famsup, paid, studytime, failures, G1, G2, G3
4. Behavioral and Lifestyle Indicatorstraveltime, freetime, goout, Dalc, Walc, health, absences

Target Variable
G3: Final grade in Mathematics (used for regression)
Pass/Fail: Derived from G3 (pass if G3 ≥ 10)

Data Preprocessing Steps: Categorical variables were encoded using Label Encoding.
Missing values were checked, but the dataset had no null entries.
Additional feature (Pass/Fail) was derived from G3 for classification tasks.

Results

Regression Task (Predicting Final Grades - G3)
Three machine learning models were trained and evaluated:

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor

Performance Comparison:

Model	R^2 Score	RMSE (Root Mean Squared Error)
Linear regression	0.69	2.48
Random forest	0.87	1.59
XGBoost	0.85	1.68

Best Model: Random Forest Regressor
Key Insight: The model can accurately predict final grades with an R^2 score of 0.87.

Reference

1. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. University of Minho, Portugal.
2. Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence.
3. Bhardwaj, B. K., & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. IJCA, Vol. 33(3).
4. Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data Mining Applications: A Comparative Study for Predicting Student's Performance. IJCA, Vol. 55(10).
5. Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology.6. UCI Machine Learning Repository. (2008). Student Performance Data Set.URL: https://archive.ics.uci.edu/ml/datasets/Student+Performance