Predicting Student Achievement Through Machine Learning: A Dual-Dataset Analysis

DMS672 Project Report

Anushri Bhargava (220189)   Meet Lalwani (220643)   Mohit Choudhary (230658)   Rohan Nimesh (220907)   Saksham Parihar (220939)

November 11, 2025

Contents

1 Introduction

Educational data analysis helps uncover key factors that impact student performance. Using statistical and machine learning techniques, we can predict academic outcomes and design data-driven strategies to support students at risk of underperforming.

This project focuses on two datasets: one representing students enrolled in Mathematics courses and the other representing students in Portuguese language courses. Each dataset contains various academic, demographic, and behavioral attributes. The central objective is to predict the final grade (`G3`) and identify the most influential factors that contribute to academic achievement.

2 Data Description

2.1 Dataset Overview

The datasets were collected by Cortez and Silva (2008) and made available via the UCI Machine Learning Repository. Both share the same structure and feature definitions, differing only by subject.

- **Mathematics dataset:** 395 student records
- **Portuguese dataset:** 649 student records

Each dataset includes 33 attributes describing a student's demographic background, family situation, study behavior, and academic results. The target variable, `G3`, represents the final grade on a scale of 0 to 20.

2.2 Feature Overview

The features fall into four main categories:

- **Demographic:** Age, gender, address type (urban/rural), family size.
- **Parental:** Education and occupation of both parents.
- **Academic:** Study time, number of past failures, absences, and periodic grades (`G1`, `G2`).
- **Social & Lifestyle:** Time spent going out with friends, free time, and Internet access at home.

No missing values were detected, and the data were consistent across both subjects.

3 Data Preprocessing

Preprocessing ensures that the dataset is clean, structured, and suitable for model training. The following steps were implemented systematically for both datasets.

3.1 Data Cleaning

- No null or duplicate values were found in either dataset.
- Text-based categorical features (e.g., "yes/no", "M/F") were standardized for consistent encoding.

- Outliers were checked in numeric variables like `absences`; extreme values were handled through capping to reduce skewness.

3.2 Feature Encoding

- Binary features (e.g., `sex`, `internet`) were converted to 0/1 representation.
- Nominal categorical variables (e.g., school, guardian type) were encoded using label encoding.
- Ordinal variables such as parental education (`Medu`, `Fedu`) retained their integer scale (0–4) to preserve the inherent order.

3.3 Feature Scaling

- Numerical variables like `studytime`, `failures`, and `absences` were normalized using Min–Max scaling to ensure all features had comparable influence.

3.4 Data Splitting

- The data were split into training (80%) and testing (20%) subsets.
- Stratified splitting ensured that the distribution of final grades remained consistent across both subsets.

These preprocessing steps established a robust foundation for subsequent model training and evaluation.

4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to gain insights into feature distributions, correlations, and relationships between independent variables and the final grade (`G3`).

4.1 General Observations

- Both datasets are balanced in terms of gender and cover a similar age range (15–22 years).
- Most students live in urban areas and have Internet access at home.
- A large proportion of students report having no prior academic failures.

4.2 Key Findings

1. Relationship between Grades: There is a very strong positive correlation between `G1`, `G2`, and `G3`. This suggests that early academic performance is a reliable indicator of final outcomes. Students who perform well initially tend to maintain good grades throughout the term.

2. Impact of Past Failures: The number of previous failures is one of the strongest negative predictors of final grades. Students with multiple past failures consistently score lower in their final results, emphasizing the long-term academic impact of repeated underperformance.

3. Effect of Parental Education: Students whose parents have higher education levels (college or above) tend to achieve higher grades. This may reflect greater academic support and motivation in households with stronger educational backgrounds.

4. Study Habits and Engagement: Higher study time correlates positively with grades, though the effect size is moderate. Conversely, frequent absences and excessive social activities (e.g., going out often) are associated with lower performance.

5. Gender and Subject Trends: In Mathematics, male students exhibit slightly higher average grades. In Portuguese, female students outperform males on average, aligning with typical trends observed in STEM and language subjects respectively.

6. Socioeconomic Influence: Students from urban areas or those with Internet access at home generally achieve better grades, indicating that access to resources plays a supportive role in learning outcomes.

4.3 Comparative Insights

While both datasets exhibit similar trends, Portuguese students tend to perform slightly better overall, with higher mean final grades and fewer failing scores. The correlation patterns are nearly identical between the two datasets, confirming that the main driving factors for performance — prior grades, parental education, and failures — remain consistent across subjects.

5 Model Building

Choice of models:-

Linear Regression | Decision Tree Regressor | Random Forest Regressor | Gradient Boosting Regressor | XGBoost Regressor

Training strategy:-

- Models were trained using a 70/30 train/test split for the leakage-free case (without G1 G2).
- For the full-feature case (with G1 and G2), 5-fold cross-validation was implemented using a preprocessing pipeline (Pipeline) to handle encoding and scaling.

Hyperparameter tuning:-

- Random Forest: n estimators=400
- Gradient Boosting: n estimators=500, learning rate=0.05
- XGBoost: n estimators=500–700, learning rate=0.05, max depth=4–6, subsample=0.8–0.9
- No grid or randomized search was applied; parameters were manually selected based on literature and experimentation.

6 Model Evaluation and Results

6.1 Portuguese Dataset

- Best Model: Random Forest (highest R^2 and lowest error metrics).
- Accuracy (R^2): 26.9
- Observation: Low predictive power due to missing grade history.
- Best Model: Random Forest
- Accuracy (R^2): 84.3
- Observation: Prediction is highly accurate when G1 and G2 are included.

Model	R^2 (%)	RMSE	MAE
Linear Regression	21.1	2.953	2.235
Decision Tree	-24.9	3.718	2.698
Random Forest	26.9	2.843	2.122
Gradient Boosting	19.8	2.978	2.290
XGBoost	26.6	2.849	2.164

Table 1: Model Performance without G1 & G2 (Test Set Results)

Model	R^2 (%)	RMSE	MAE
Random Forest	84.3	1.268	0.824
Linear Regression	83.6	1.297	0.844
XGBoost	83.5	1.297	0.834
Gradient Boosting	83.2	1.310	0.839
Decision Tree	69.1	1.773	1.018

Table 2: Model Performance with G1 & G2 using 5-Fold Cross-Validation

6.2 Mathematics Dataset

Model	R^2 (%)	RMSE	MAE
Linear Regression	19.1	4.218	3.288
Decision Tree	~ 0 or negative	—	—
Random Forest	28.9	3.954	3.115

Table 3: Model Performance for Mathematics Dataset without G1 & G2

- Best Model: Random Forest
- Accuracy (R^2): 28.9
- Observation: Final grade prediction is poor without prior grade context.

Model	R^2 (%)	RMSE	MAE
Random Forest	87.0	1.599	1.023
XGBoost	86.9	1.603	1.074
Gradient Boosting	86.4	1.639	1.049
Linear Regression	79.2	2.057	1.378
Decision Tree	76.4	2.105	1.132

Table 4: Model Performance for Mathematics Dataset with G1 & G2 (5-Fold Cross-Validation)

- Best Model: Random Forest
- Accuracy (R^2): 87.0
- Observation: Predictive accuracy is very high with G1 and G2 included.

7 Result Interpretation

7.1 Portuguese Dataset

Feature Importance

Failures, study time, absences, parental education, and alcohol consumption showed the most influence in the leakage-free model. When G1 and G2 were included, they dominated the feature importance, explaining approximately 80% of the prediction variance.

Predicted vs Actual

- **Without G1/G2:** Wide scatter and poor alignment between predicted and actual grades.
- **With G1/G2:** Strong linear trend along the $y = x$ line with low residual errors.

Educational Implications

Early-term grades (G1 and G2) are critical indicators of final outcomes. Intervention efforts must begin early, focusing on improving attendance, reducing the likelihood of failures, and providing enhanced support for students who show early signs of academic risk.

Strategies to Enhance Performance

- Early detection of low G1/G2 scores for timely academic intervention.
- Implementing programs to reduce student absences and increase study time.
- Providing family and institutional support to students from at-risk demographics.

7.2 Mathematics Dataset

Feature Importance

Feature importance patterns were similar to the Portuguese dataset. However, math-specific variables such as absences, failures, and study time had slightly stronger influence. When G1 and G2 were included, these two features again overwhelmingly dominated the prediction process.

Predicted vs Actual

- **Without G1/G2:** High residuals and poor predictive fit.
- **With G1/G2:** Clear diagonal alignment between predicted and actual values, indicating highly accurate predictions.

Educational Implications

Mathematical performance closely tracks prior-term grades. Without G1 and G2, demographic and lifestyle variables provide weak predictive signals for final performance.

Strategies

- Use G1 and G2 as early indicators for identifying students needing support.
- Provide targeted instructional reinforcement immediately after G1 results.
- Monitor and reduce absenteeism, particularly in coursework with heavy mathematical content.

8 Conclusion

The analysis of the Portuguese and Mathematics datasets shows that academic outcomes are shaped by behavioral, familial, and academic factors, with study time, failures, absences, and parental background showing meaningful influence. However, models trained without prior grades (G1 and G2) exhibited weak predictive power, with the best R^2 values below 30%. When G1 and G2 were included, model accuracy rose sharply, reaching over 84% for Portuguese and 87% for Mathematics, confirming early-term grades as the strongest predictors of final performance. Overall, the findings highlight that while demographic and lifestyle factors alone offer limited predictive value, incorporating early academic assessments enables highly accurate identification of at-risk students and supports timely educational intervention.

9 Contribution

- Meet Lalwani : EDA code, Modelling, Report Writing(Results & Interpretation)
- Mohit Choudhary : EDA code, Modelling, Report Writing(Model Encoding & Scaling)
- Anushri Bhargava : EDA code, Modelling, Report Writing(Data Preprocessing)
- Rohan Nimesh : EDA code, Modelling, Report Writing(Model Evaluation & Results)
- Saksham Parihar : EDA code, Modelling, Report Writing(EDA & Description)

10 Acknowledgments

The datasets used in this project were obtained from the UCI Machine Learning Repository: *"Student Performance Data Set"* (Cortez & Silva, 2008).