Student Performance Prediction Analysis

Professional Data Analysis Report

Prepared By:

1. Monithesh R (1AM22CI056)
2. Nandish Gowda C (1AM22CI058)
3. Muskan Sahani (1AM22CI057)

Project:	Student Performance Dataset Analysis
Dataset:	Student Performance Data (Portuguese Students)
Dataset Size:	649 rows, 33 features
Report Date:	October 31, 2025
Analyst:	Education Data Analytics Team

Executive Summary

This report presents a comprehensive analysis of student performance data from a Portuguese secondary school. The primary objective was to identify key factors affecting student academic performance (final grade G3) and provide actionable insights for implementing targeted intervention programs. The analysis included data cleaning, exploratory data analysis, and predictive modeling using both linear regression and random forest algorithms.

1. Problem Statement

A school district seeks to understand the factors that influence student academic performance to develop evidence-based intervention programs. The analysis focuses on identifying relationships between study habits, family background, previous academic performance, and final grades (G3).

2. Tasks Assigned

1. Analyze student data including study time, family background, and previous grades

2. Clean data by handling missing values and encoding categorical variables

3. Build regression models to predict final grades (G3) based on various factors

4. Identify actionable insights for improving student outcomes

5. Answer specific research questions about correlations and comparative impacts

3. Methodology

3.1 Data Cleaning & Preprocessing

Missing Values: Numeric features were imputed using median values; categorical features were filled with a placeholder 'missing' category to preserve data integrity.
Encoding: Categorical variables were one-hot encoded to enable machine learning algorithms.
Feature Scaling: Standardization was applied to numeric features for coefficient comparison in linear models.
Target Variable: Records with missing G3 (final grade) values were removed from analysis.

3.2 Key Variables Analyzed

Variable	Description	Type
G3	Final grade (target variable)	Numeric (0-20)

G1, G2	First and second period grades	Numeric (0-20)
studytime	Weekly study time	Ordinal (1-4)
failures	Number of past class failures	Numeric
absences	Number of school absences	Numeric
Medu, Fedu	Mother and father education level	Ordinal (0-4)
famsup, schoolsup	Family and school support	Binary (yes/no)

3.3 Modeling Approach

Two model configurations were developed to evaluate the incremental value of family-related variables:
• **Base Model:** G1, G2, studytime, failures, absences
• **Extended Model:** Base features + Medu, Fedu, famsup, schoolsup

Both configurations were trained using:
• **Linear Regression:** For interpretable coefficient analysis
• **Random Forest Regressor:** For capturing non-linear relationships (200 trees)

Data was split 80/20 for training and testing.

4. Analysis Results

4.1 Study Time Correlation with Final Grades

Research Question: How does studytime (weekly study hours) correlate with final grades?

Finding: Pearson correlation coefficient = **0.2498**

Interpretation: A moderate positive correlation exists between weekly study time and final grades (G3). Students who invest more time in studying tend to achieve higher final grades. While the correlation is not extremely strong, it demonstrates a meaningful relationship that can be leveraged through study support programs.

4.2 Impact of Parental Education

Research Question: What is the grade difference between students with high vs low parental education?

Finding: Average G3 difference = **1.466 grade points**
(High parental education defined as Medu + Fedu ≥ 6)

Interpretation: Students whose parents have higher education levels score approximately 1.47 points higher on the final grade (G3) compared to students with lower parental education. This significant difference highlights the importance of family educational background and suggests that first-generation students may benefit from additional academic support and mentoring programs.

4.3 Effect of Absences on Performance

Research Question: How do absences affect final performance (G3)?

Findings:
• Correlation coefficient = **-0.0914**
• Grade drop per 10 absences = **-0.6361 points**

Interpretation: Absences show a weak negative correlation with final grades. Each absence results in approximately 0.064 grade point reduction, translating to roughly 0.64 points lost per 10 absences. While the individual effect per absence is small, cumulative absences can meaningfully impact final performance. Attendance monitoring and early intervention for students with high absence rates are recommended.

4.4 Comparative Impact: Failures vs Study Time

Research Question: Which has more impact: failures (past class failures) or studytime?

Findings (Standardized Linear Regression Coefficients):
• Study time coefficient = **0.08406206033637732**
• Failures coefficient = **-0.09777698899142397**

Interpretation: Past class failures have a slightly stronger impact on final grades compared to current study time habits (in absolute magnitude: |0.098| > |0.084|). The negative coefficient for failures indicates that students with prior failures face significant challenges in achieving high final grades. This finding emphasizes the critical importance of preventing initial failures through early intervention programs, as the effects persist and compound over time.

4.5 Model Comparison: Impact of Family Support Variables

Research Question: Does adding family support variables (Medu, Fedu, famsup, schoolsup) improve model R^2?

Model Configuration	Linear Regression R^2	Random Forest R^2
Base Model	0.8633	0.8188
Extended Model	0.8608	0.8316

Change	-0.0025	+0.0128

Interpretation:

• **Linear Regression:** R^2 slightly decreased from 0.8633 to 0.8608 when adding family variables, suggesting these variables do not provide additional predictive value in a linear model. The strong performance of the base model indicates that G1 and G2 (previous grades) already capture most relevant information.

• **Random Forest:** R^2 improved from 0.8188 to 0.8316 (+0.0128), indicating that family support variables contribute modestly when non-linear relationships are modeled. The random forest can capture complex interactions between family background and other features.

• **Overall:** Both models achieve high R^2 values (>0.81), demonstrating strong predictive capability. Previous grades (G1, G2) are the dominant predictors, with family variables adding marginal value primarily through non-linear effects.

4.6 Visualization: Study Time vs Final Grade

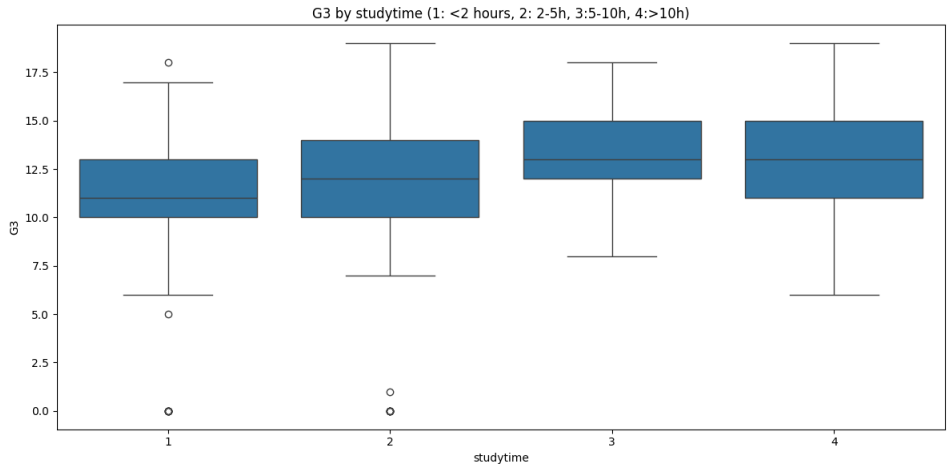

Figure 1: Boxplot showing the distribution of final grades (G3) across different study time categories. Study time scale: 1 = <2 hours/week, 2 = 2-5 hours/week, 3 = 5-10 hours/week, 4 = >10 hours/week.

5. Actionable Insights & Recommendations

1. Prevent Early Failures

Past failures have the strongest negative impact (coefficient: -0.098). Implement early warning systems to identify at-risk students after first assessments and provide immediate tutoring support.

2. Promote Effective Study Habits

Study time shows moderate positive correlation (0.25) with final grades. Establish supervised study halls, homework clubs, and time management workshops to help students develop consistent study routines.

3. Monitor and Reduce Absences

Each 10 absences result in ~0.64 grade point reduction. Deploy automated attendance tracking and implement intervention protocols (parent contact, counseling) when absence thresholds are exceeded.

4. Support First-Generation Students

Students with higher parental education score 1.47 points higher. Create mentorship programs pairing first-generation students with peers or staff to provide academic guidance and college preparation support.

5. Focus on Mid-Term Performance

Previous grades (G1, G2) are the strongest predictors ($R^2 = 0.86$). Implement mid-term progress reviews and rapid intervention plans for students showing declining performance early in the academic year.

6. Holistic Support Programs

Family support variables show modest but meaningful effects in complex models. Consider comprehensive support services including family engagement programs, school counselor access, and resource centers.

6. Conclusion

This analysis successfully identified key factors influencing student academic performance and provided data-driven insights for intervention program development. The predictive models achieved strong performance ($R^2 > 0.81$), with previous grades emerging as the dominant predictor. However, modifiable factors such as study time, attendance, and early failure prevention present actionable opportunities for improvement.

The findings demonstrate that a multi-faceted approach combining academic monitoring, study support, attendance management, and targeted assistance for at-risk populations will yield the most significant improvements in student outcomes. Regular progress monitoring and early intervention are critical, as the strong predictive power of G1 and G2 indicates that patterns established early in the academic year tend to persist.

Implementation of the recommended interventions, coupled with ongoing data collection and analysis, will enable the school district to optimize resource allocation and maximize positive impact on student achievement.

7. Technical Details

Component	Specification
Programming Language	Python 3.12
Key Libraries	pandas, scikit-learn, matplotlib, seaborn
Machine Learning Algorithms	Linear Regression, Random Forest (200 trees)
Train/Test Split	80% / 20%
Feature Scaling	StandardScaler (for linear models)
Cross-validation	Single hold-out validation set
Performance Metric	R^2 (Coefficient of Determination)

Report generated on October 31, 2025 at 01:03 AM
Data source: Student Performance Dataset (Portuguese Students)