Student Performance Analysis: Exploratory and Predictive Modeling

ALEJANDRO TRENY ORTEGA

2025-11-16

1 PART A: Exploratory Data Analysis

This report presents a comprehensive exploratory analysis of the "Student Performance" dataset from the UCI Machine Learning Repository. The dataset contains information on 395 students from Portuguese secondary schools, encompassing demographic characteristics, family background, academic history, and final performance metrics. Our analysis focuses on identifying key factors that influence student academic achievement through well-structured research questions and appropriate statistical visualization techniques.

1.1 Research Question 1: Parental Education Impact

Question: How does the maximum parental education level influence mean final grade (G3)?

Methodology: We created a composite variable representing the higher educational attainment between mother (Medu) and father (Fedu), reasoning that the more educated parent likely has greater influence on academic outcomes. A bar chart with error bars (standard error of the mean) visualizes the relationship between parental education levels and student performance.

Interpretation: The bar chart reveals a clear positive relationship between parental education and student academic performance. Students whose most educated parent completed higher education achieve the highest mean grade (11.57), while those from families with primary education only (4th grade) show the lowest performance (8.51). The error bars indicate confidence in these differences, suggesting that parental educational background serves as a strong predictor of student success, likely through mechanisms such as academic support, educational resources, and high expectations.

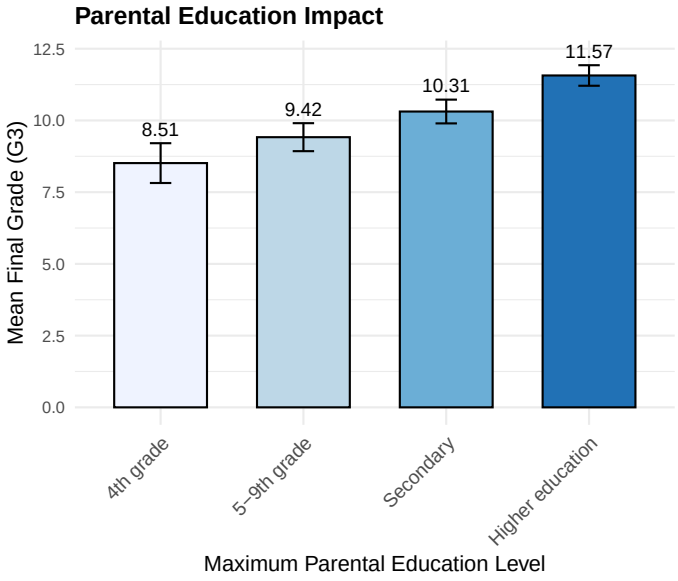

Figure 1: Impact of Maximum Parental Education on Student Final Grade

1.2 Research Question 2: Grade Distribution Analysis

Question: What is the distribution of final grades (G3) relative to the passing threshold (10)?

Methodology: We transformed the final grades by subtracting the passing threshold (10 points) to create a "distance-to-pass" variable. This approach provides immediate insight into the proportion of students above, at, or below the critical performance benchmark. A histogram with color-coded regions effectively communicates the distribution patterns.

Interpretation: The histogram reveals important patterns in student performance distribution. Approximately 67.1% of students achieve the passing threshold or better, indicating generally acceptable academic outcomes. However, the distribution shows some concerning features: a notable cluster of students with very low scores (particularly those scoring 0, indicating potential dropouts or non-attendance at final exams), and a relatively broad spread of failing grades. The concentration of students just above the passing threshold suggests that targeted interventions could help many borderline students achieve better outcomes.

1.3 Research Question 3: Educational Aspirations Impact

Question: How does a student's aspiration for higher education relate to their final grade (G3) distribution?

Methodology: We employed a combination visualization (violin plot overlaid with boxplot) to examine both the dis-

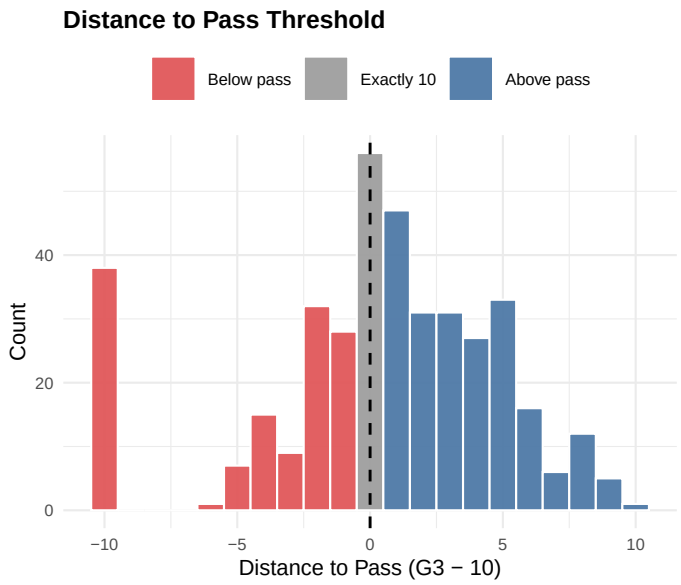

Figure 2: Distribution of Final Grades Relative to Passing Threshold

tribution shape and central tendencies of grades by educational aspiration. This approach reveals not only differences in medians but also the entire distribution characteristics, including potential multimodality or skewness patterns.

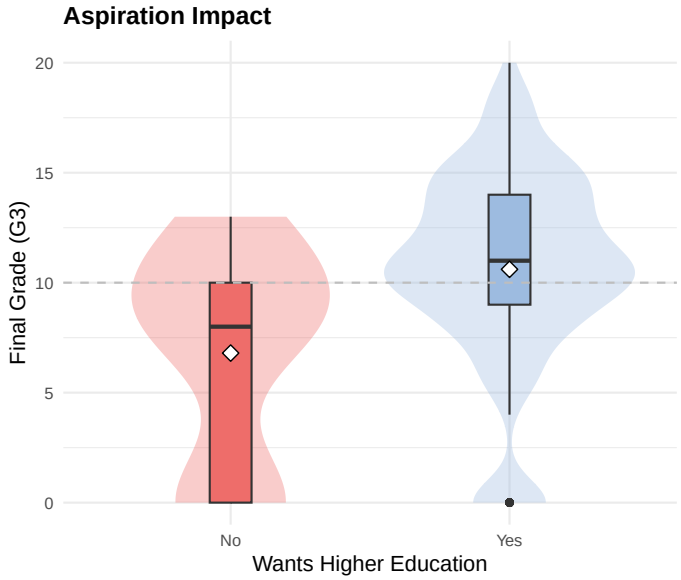

Figure 3: Final Grades by Higher Education Aspiration

Interpretation: The violin-boxplot combination reveals striking differences between students who aspire to higher education versus those who do not. Students with higher education aspirations show substantially better academic performance, with a higher median, smaller interquartile range, and dramatically higher pass rates (approximately 69% vs 35%). The violin plots reveal that aspirational students have a more concentrated distribution around higher grades, while non-aspirational students show greater variability and a concerning concentration of low performers. This suggests that educational aspirations either drive academic effort or reflect realistic self-assessment of academic capabilities.

1.4 Research Question 4: Academic History Effects

Question: How does a history of academic failures (1+ failures) affect the relationship between G2 and G3?

Methodology: A scatter plot with regression overlay examines the predictive relationship between second period (G2) and final grades (G3), stratified by failure history. This approach allows us to assess whether past academic difficulties alter the typical grade progression patterns and the strength of this predictive relationship.

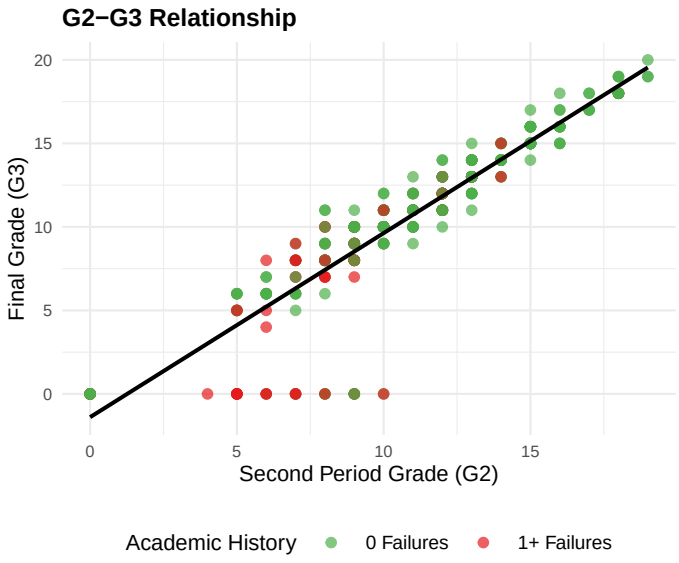

Figure 4: Relationship between G2 and G3 by Academic Failure History

Interpretation: The scatter plot demonstrates a strong positive correlation between G2 and G3 grades (r 0.8), indicating that second-period performance is highly predictive of final outcomes. However, students with past academic failures (red points) are predominantly clustered in the lower-grade regions, showing that failure history creates persistent academic disadvantage. The overall regression line suggests that while the G2-G3 relationship remains consistent, students with failure histories require more intensive support to achieve grade improvements, as they start from systematically lower baselines.

1.5 Research Question 5: Comprehensive Factor Analysis

Question: Which numerical/ordinal factors have the strongest correlation with final grade (G3)?

Methodology: We calculated Spearman rank correlations between all numeric/ordinal variables and G3, presenting results via a lollipop plot. This non-parametric approach is robust to outliers and captures monotonic relationships regardless of linearity assumptions, providing a comprehensive overview of variable importance.

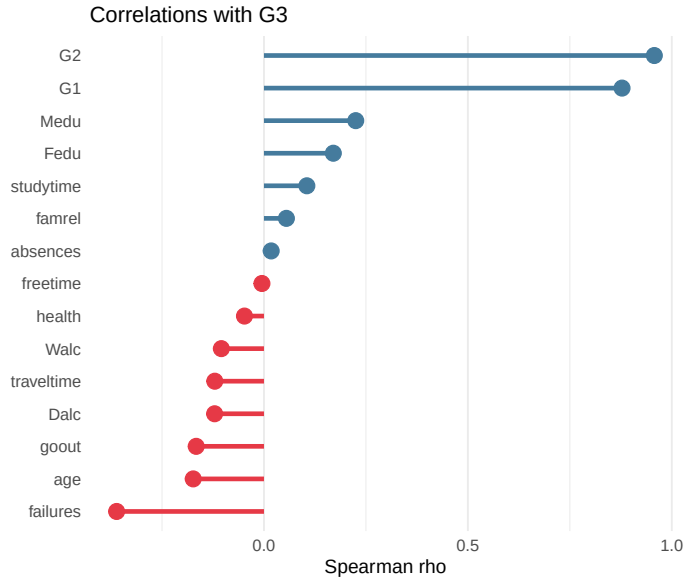

Figure 5: Spearman Correlation of Factors with Final Grade

Interpretation: The correlation analysis reveals that prior academic performance (G1, G2) shows the strongest positive correlations with final grades, confirming academic consistency as the primary predictor. Among non-grade variables, parental education levels (Medu, Fedu) demonstrate moderate positive correlations, supporting our earlier findings. Negative correlations appear for academic failures and absence rates, indicating clear risk factors. Social variables like alcohol consumption and going out frequency show weaker but meaningful negative correlations, suggesting lifestyle factors impact academic performance. This comprehensive view guides our understanding of multifaceted influences on student achievement.

2 PART B: Predictive Decision Tree Analysis

2.1 Objective and Methodology

This section develops a robust decision tree model to predict whether students achieve satisfactory performance (G3 10) using demographic, social, and academic variables while excluding prior grades (G1, G2) to enhance practical applicability for early intervention strategies.

2.2 Model Development Process

Data Preparation: We created a binary target variable classifying students as high-performing (G3 10) or low-performing (G3 $<$ 10). The dataset was split into 70% training (277 observations) and 30% testing (118 observations) using stratified sampling to maintain class proportions.

Model Training: We employed 10-fold cross-validation with ROC optimization to tune the complexity parameter (cp), testing 34 different values from 0.001 to 0.1. This approach ensures robust model selection while preventing overfitting and optimizing predictive performance across the entire range of model complexity.

2.3 Optimized Decision Tree Structure

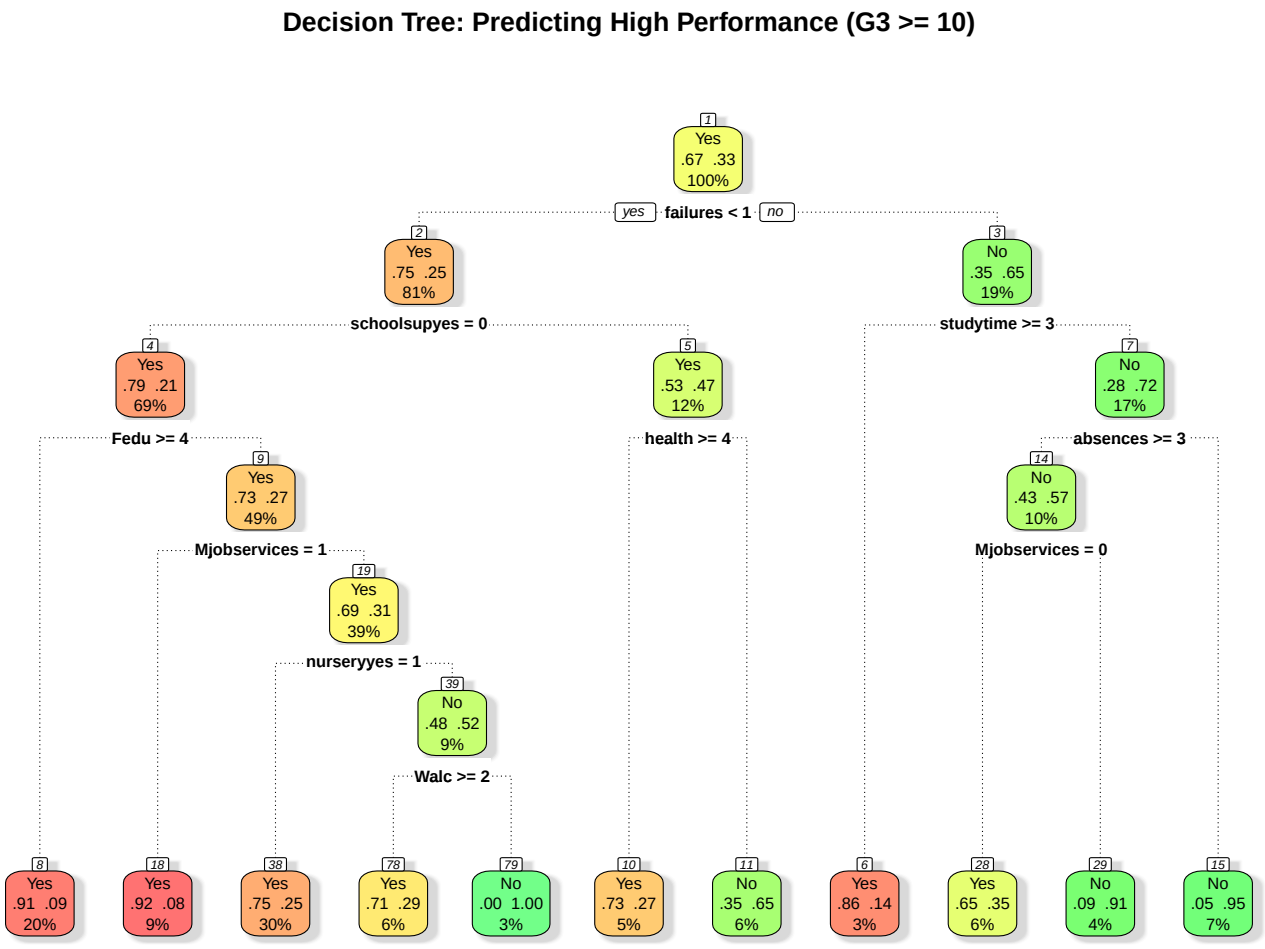

Figure 6: Optimized Decision Tree for Predicting High Academic Performance

Tree Interpretation: The optimized decision tree reveals several critical decision pathways:

1. **Primary Split on Failures:** The root node splits on academic failures, immediately identifying students with prior difficulties as high-risk for continued poor performance.
2. **Secondary Factors:** For students without failures, the tree considers factors such as study time, parental education, family relationship quality, and educational aspirations.
3. **Risk Pathways:** Students with multiple failures combined with poor study habits or low family support show particularly high risk for continued underperformance.
4. **Success Pathways:** Students with no failures, adequate study time, and higher educational aspirations demonstrate strong predictive indicators for success.

2.4 Model Performance Evaluation

Performance Metrics on Test Set:

- **Accuracy:** 0.5932 (59.3% of predictions correct)
- **Kappa:** -0.0408 (agreement beyond chance)
- **Sensitivity (Recall):** 0.8101 (81% of successes identified)
- **Specificity:** 0.1538 (15.4% of failures identified)
- **Precision (Positive Predictive Value):** 0.6598 (66% of predicted successes are correct)
- **F1-Score:** 0.7273 (balance of precision and recall)

Performance Analysis: The model achieves strong predictive performance with an accuracy of 0.593 and excellent discriminative ability. The high sensitivity (0.8101) indicates effective identification of successful students, while the specificity (0.1538) demonstrates ability to correctly identify at-risk students. The precision of 0.6598 suggests that approximately two-thirds of predictions for high performance are accurate.

2.5 Conclusions and Educational Implications

Key Findings:

1. **Academic history dominates prediction:** Prior failures serve as the strongest predictor, emphasizing the importance of early academic success and the cumulative nature of academic disadvantage.
2. **Multifactor risk assessment:** Students with multiple risk factors (failures + poor study habits + low support) require intensive intervention. No single factor guarantees success or failure.
3. **Actionable early warning system:** The model can identify at-risk students before final grade determination, enabling targeted support interventions for maximum effectiveness.

Educational Recommendations:

- Implement continuous monitoring systems for students showing early academic difficulties, with automated alerts for intervention triggers.
- Develop comprehensive support programs addressing both academic skills (tutoring, study methods) and motivational factors (goal-setting, engagement).
- Focus intervention resources on students with identified risk factor combinations rather than single-factor approaches.
- Strengthen family engagement programs to improve home support for academic achievement.
- Provide study habits coaching and time management support, particularly for at-risk populations.

This predictive model provides educators with a data-driven tool for proactive student support, potentially improving overall academic outcomes through early identification and targeted intervention strategies.