Executive Summary:
Machine Learning Models for Student Grades

Author: Steph Roberts

Executive Summary

This project aimed to analyze student data to develop a machine learning model for use in predicting student performance. By analyzing key features such as past grades, absences, and study habits, we identified critical factors that influence student outcomes. Our analysis revealed that previous grades are the most significant predictors of future performance, while absences and past failures negatively impact grades. Based on these findings, we recommend targeted interventions for students with frequent absences or past failures. Additionally, we suggest offering in-school study time to support students by increasing their overall study time. This could be particularly beneficial for students with long travel times, those who have had any failures, or those who are prone to absences, as it can help them catch up on missed curriculum and improve their academic performance.

Problem Framing & Big Picture

The primary objective of this project was to create a predictive model that could identify students at risk of underperforming, allowing the Advising Team to intervene proactively. The model's outcomes can lead to actionable decisions, such as identifying students who require additional academic support or structuring personalized study plans. By using this model, the school system can provide targeted interventions to improve students' grades, ultimately enhancing educational outcomes and reducing the likelihood of students falling behind.

Recommendations

1. **Intervene after Any Failure:** Given the moderate negative correlation between past failures and final grades, it is crucial to intervene as soon as a student experiences a failure to prevent further academic decline.
2. **Address Absenteeism:** Students with more than 5 absences in a term should be flagged for intervention to mitigate the negative impact on their grades.
3. **Implement In-School Study Periods:** Offering an elective class period dedicated to in-school study time could benefit students with long travel times and limited opportunities to study at home, thereby enhancing their academic performance.
4. **Enhance School Pride:** For students attending the school primarily due to proximity, initiatives to boost school pride, such as pep rallies and team-building events, could foster a stronger connection to the school and positively influence their performance.

Data Overview

The dataset used for this analysis includes student performance data from two Portuguese schools, with key attributes such as past grades (G1, G2), absences, failures, study time, and more. The target variable is the grades for the third term (G3). Given that school leaders may need to implement interventions before the grades from the first two terms (G1 and G2) are available, we created two models: one that includes these grades and one that does not.

The model with G1 and G2 grades is significantly better at predicting future grades, as these grades are strong indicators of a student's performance. However, we also developed a model without these grades to provide insights when early-term grades are not yet available. Even without the grades, the model can identify students who might be at risk of underperforming based on other factors such as absences and failures. This allows school leaders to intervene early, offering support such as in-school study periods to help students catch up on missed or misunderstood curriculum, potentially improving their final grades. This dual-model approach ensures flexibility, enabling the system to make informed decisions even when full grade data isn't available, making it a valuable tool for early intervention strategies.

Analytical Insight

The key findings from our comprehensive data exploration reveal the most significant factors influencing student performance. A variety of analytical techniques, including statistical analysis and visualizations, were employed to uncover patterns and correlations. These insights highlight the drivers of academic success and provide a foundation for recommendations and model development.

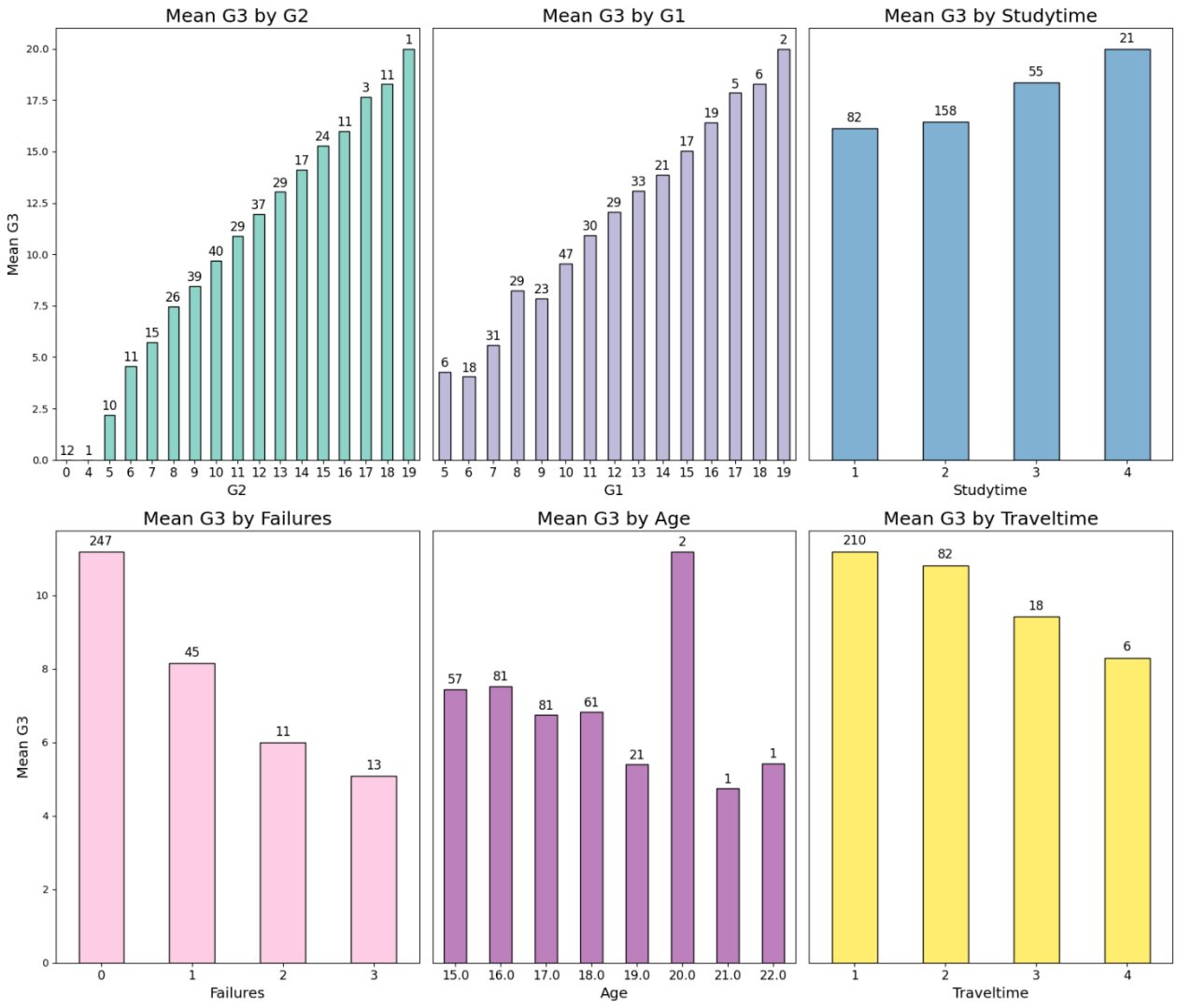

Figure 1. Bar graphs showing the relationship between key numeric variables and final term grades (G3).

The bar graphs for G1 and G2 reveal a strong positive correlation with the final grades (G3), confirming that previous academic performance is a significant predictor of future success. The study time graph similarly shows that students who dedicate more time to studying tend to achieve higher final grades, reinforcing the importance of in-school study periods to enhance overall performance. Conversely, the graph for failures highlights a clear negative correlation with G3, indicating that students with more past failures generally have lower final grades, underscoring the need for early intervention. The age graph suggests that older students tend to have slightly lower grades, possibly due to additional responsibilities such as part-time jobs, making in-school study periods particularly beneficial for them. Lastly, the travel time graph demonstrates a weak negative correlation with grades, suggesting that while the impact is mild,

longer commutes can slightly hinder academic performance, which could be mitigated by providing extra study time at school.

The scatter plot (Figure 2.) reveals that as absences in the third term increase, students' final grades (G3) tend to decrease significantly. The data suggests that the impact of absences becomes particularly pronounced at around 14 absences, where grades sharply decline. To address this, it's recommended that interventions start earlier, around 5-6 absences, to prevent students from reaching this critical threshold. Proactive monitoring, tailored support for at-risk students, and continuous tracking of absences can help mitigate the negative effects of absenteeism. Educating students and parents on the importance of regular attendance is also essential to improve academic outcomes.

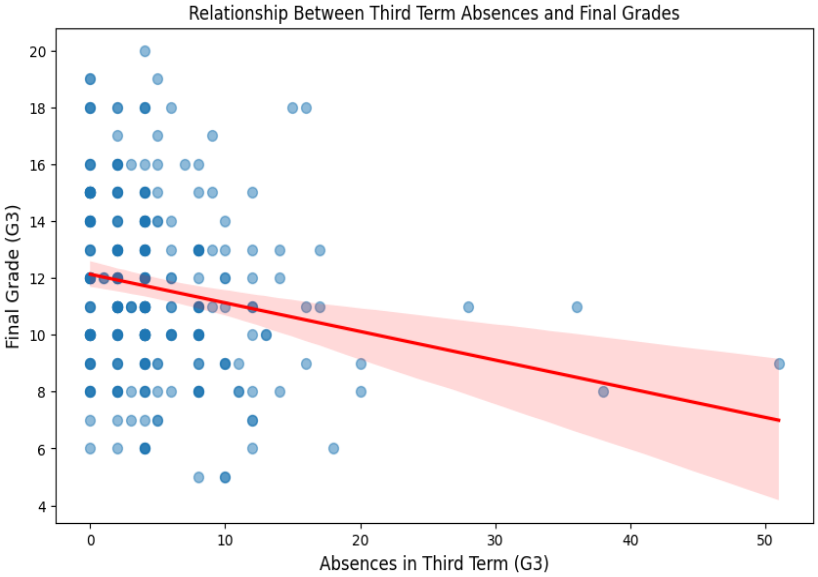

Figure 2. Scatterplot comparing absences to grades in the 3rd term.

The bar plots (Figure 3) provide a visual comparison of the mean final grades (G3) across different categorical variables, offering insights into how various factors influence academic performance. Notably, students who selected the school based on its reputation or other reasons tend to achieve higher final grades compared to those who chose it for proximity to home or specific courses. This suggests that students attending the school for its reputation may be more motivated or have a stronger commitment to their education. This insight supports the recommendation to increase school pride through initiatives like pep rallies, team-building events, and marketing efforts. By fostering a sense of pride and belonging,

students who chose the school for its proximity might be encouraged to care more about their academic performance.

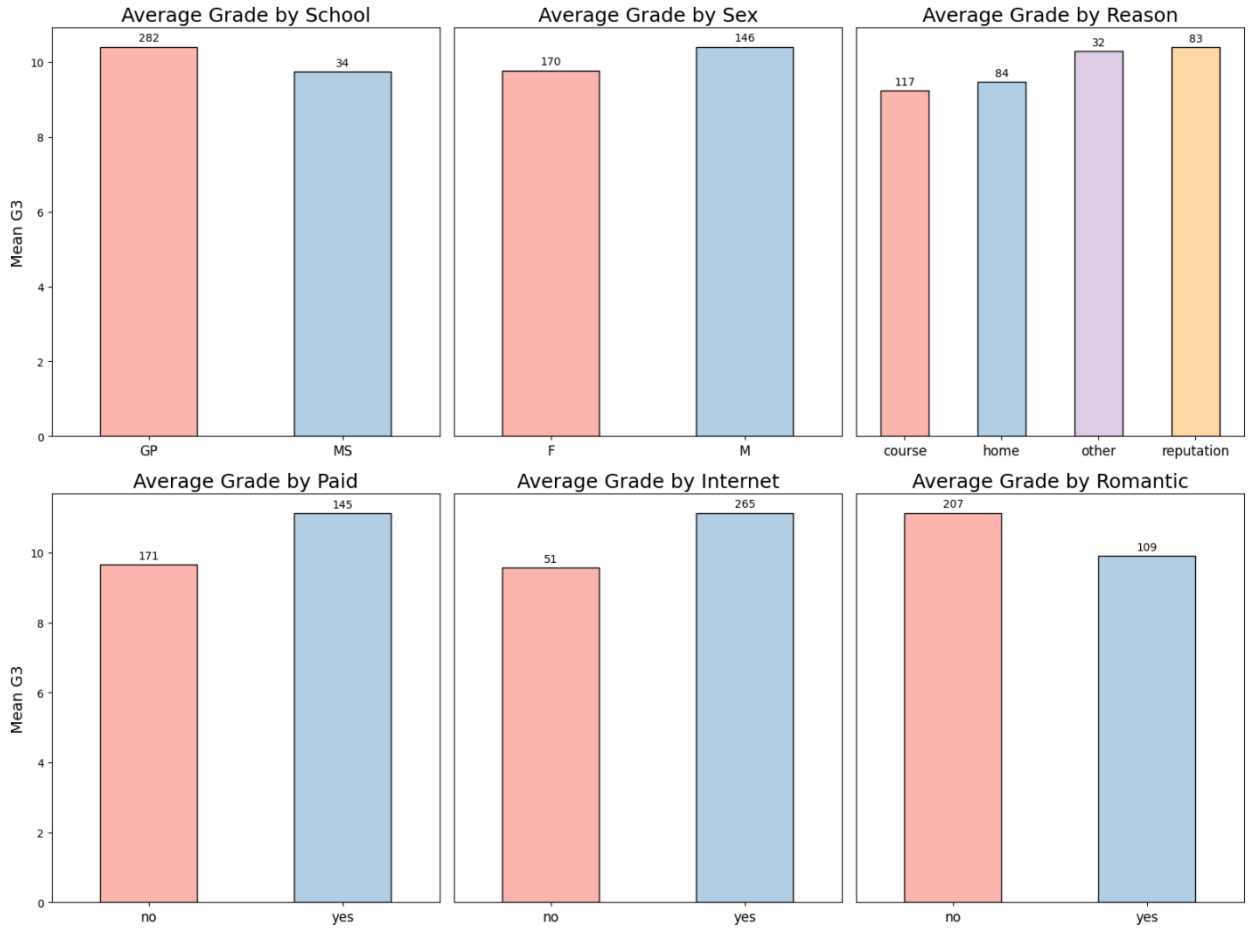

Figure 3. Bar plots showing several categorical variables grouped and averaged for 3rd term grades.

In addition, the plots show that students not in a romantic relationship generally achieve slightly higher grades compared to those who are. While romantic relationships are a natural part of teenage life and can't be controlled, this finding highlights the importance of providing structured in-school study time. By offering additional study periods, the school can help students manage their time better and ensure they have sufficient time to focus on their academic responsibilities, particularly for those whose time outside of school may be impacted by relationships or other commitments.

The box plot (Figure 4) illustrates the distribution of final grades (G3) across different levels of past failures. As the number of failures increases, the median and overall distribution of grades shift downward, indicating a clear negative impact on academic performance. This visual reinforces the finding that past failures are strongly correlated with lower grades, highlighting the importance of early interventions to prevent further academic decline. The sharp drop in grades among students with multiple failures emphasizes the critical need for targeted support to help these students improve their academic outcomes.

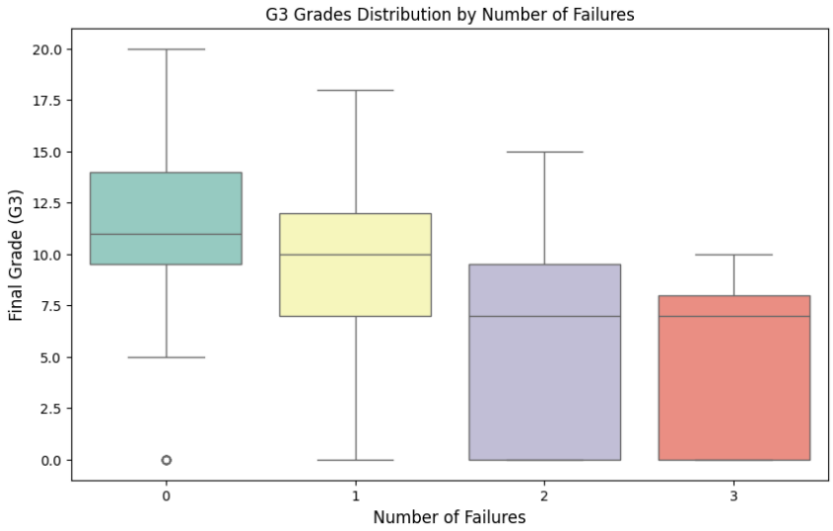

Figure 4. Box plot of failures and average term 3 grade.

The data analysis identified past grades, study time, absences, failures, reasons for school choice, and personal relationships as the most influential factors affecting student performance. By incorporating these key factors, the models are designed to effectively support school leaders in making data-driven decisions that enhance student outcomes.

Methodology

To predict student performance, we explored three different machine learning models: Linear Regression, Support Vector Machine (SVM) Regression, and Lasso Regression. Each of these models was evaluated based on their ability to predict the final grades (G3) using different sets of input variables. Before training the models, the original dataset was split into **80% training data** and **20% test data**. This split is crucial for building and validating the model. The training data is used to develop the model by identifying patterns and relationships in the data, while the test data is reserved to evaluate how well the model performs on unseen data. This helps ensure that the model can generalize to new, real-world data and is not simply memorizing the training examples.

Linear Regression is a model that finds a straight-line relationship between the input variables and the target variable, making it highly effective when there is a strong linear correlation in the data. This model was particularly relevant when we included G1 and G2 grades in the data, as these earlier grades are strong indicators of final performance.

SVM Regression is a more sophisticated model that can capture complex, non-linear relationships in the data. It works by finding the optimal boundary that best separates the data points. To optimize the SVM model's performance, we conducted a grid search to fine-tune its hyperparameters. This model showed strong performance, particularly when G1 and G2 grades were not included, as it was able to identify patterns in the data that were not captured by simpler models.

Lasso Regression is similar to Linear Regression but includes a penalty that helps reduce the impact of less important variables, potentially preventing overfitting. However, in our analysis, Lasso Regression did not outperform the other models and was therefore not chosen as the final model.

To measure the performance of these models, we used **Root Mean Squared Error (RMSE)** and **R^2 (R-squared)**. RMSE measures the average magnitude of the errors in the model's predictions, with lower

values indicating more accurate predictions. R^2 measures how well the model explains the variability in the target variable, with higher values indicating a better fit. The following are the RMSE values for each model on the training data (remember, a lower RMSE indicates a stronger model):

- **Linear Regression with Grades:** 1.8904
- **Linear Regression without Grades:** 4.2891
- **SVM with Grades:** 2.5320
- **SVM without Grades:** 4.2765
- **Lasso Regression with Grades:** 2.1691
- **Lasso Regression without Grades:** 4.3865
- **Tuned SVM with Grades:** 1.7127
- **Tuned SVM without Grades:** 4.1756

While the tuned SVM model demonstrated slightly better RMSE (1.71) on the training data with grades, the Linear Regression model had the second best RMSE (1.89) and was ultimately chosen for its simplicity and its strong linear relationship with the data. Given the very strong correlation between earlier grades and final grades, Linear Regression was deemed more likely to generalize well to new students, minimizing the risk of overfitting.

Key Results

When evaluated on the test data, the Linear Regression model with grades resulted in an RMSE of 2.2224 and an R^2 of 0.7591, indicating a strong fit and reliable prediction accuracy. Conversely, the tuned SVM model without grades had an RMSE of 4.1289 and a much lower R^2 of 0.1686.

For data without grades, the tuned SVM model was the preferred choice due to its ability to handle the more complex relationships in the absence of strong predictors like G1 and G2. However, it's important to note that the low R^2 value of 0.1686 for the model without grades indicates that this model does not explain much of the variability in final grades, which means it is not as reliable as the model with grades. While the SVM model without grades can still provide guidance to school leaders, any previous grades from a student would be far more predictive and preferred for making intervention decisions. When such grades are unavailable, this model can serve as a supplementary tool to identify students who may benefit from additional support.

Based on these results, we recommend implementing the Linear Regression model for scenarios where G1/G2 grades are available. If these grades are not accessible, the SVM model provides a viable alternative, although it should be used with caution given its lower predictive accuracy.

Conclusion

This project successfully identified key factors influencing student performance and demonstrated the value of using past grades in predictive models. The linear regression model proved to be the most reliable when prior grades are available, offering accurate predictions that can inform timely interventions. This result highlights the strong linear relationship between earlier and final grades, making this model particularly effective for school leaders seeking to identify students who may need support.

However, challenges remain in predicting performance without these grades, as evidenced by the lower accuracy of the SVM model when G1 and G2 grades were not included. The poor performance of the model without grades suggests that additional data could be crucial for improving predictions in these cases. One potential next step is to continue gathering data for students each term or academic year and reevaluate the model's performance. With more data instances, this model—or potentially a different one—could be better equipped to predict grades and outcomes even without prior grades.

Additionally, it would be beneficial to investigate specific instances, such as students who received a grade of 0 (see Figure 5). These cases are unusual and could indicate circumstances like switching schools, dropping out, or other unique situations. Understanding these instances better could help determine if they are true outliers or if they reveal patterns that should be accounted for in the model. This deeper analysis could refine the model further, ensuring it accurately reflects the real-world scenarios it is meant to predict.

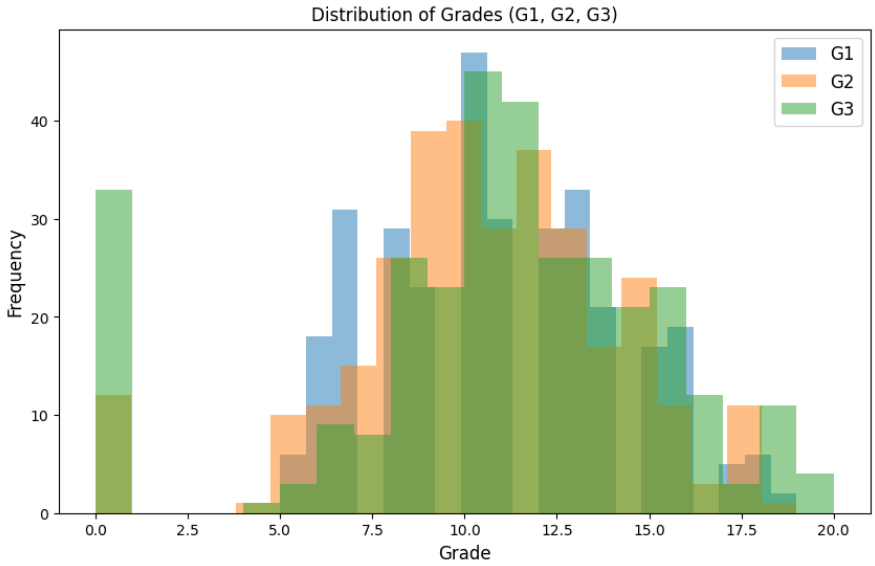

Figure 5. Overall distribution of grades by term.

Overall, while the project achieved its primary objectives, the findings indicate that further refinement and data collection could enhance the accuracy and reliability of predictive models, particularly in scenarios where early-term grades are unavailable. Investigating anomalies and specific cases could provide additional insights, leading to even more robust and actionable predictions.