STOR 320: Introduction to Data Science

Submit to gradescope by 11:59 PM, Dec 09, 11:59 PM.

Final Paper Group 13

Group members: Semeon Petros, Lu Wang, Helen Fu, Meihan Chen

Introduction

Parents and educators have a deep concern for potential factors that may impact a student's academic performance. Intellectual ability plays an undeniable part in academic success, but the growing library of research points to the important role of environmental influences as well. Of these factors, little, if any, attention has been given to the family dynamics and atmosphere that may affect grades. In this research, we investigate how various family-related factors affect student performance in Mathematics and Portuguese, the two core subjects in Portugal. As part of our analysis, we use data collected by Professor Paulo Cortez from two public high schools in Portugal during the 2005-2006 academic year. We seek to examine two central questions that provide new insight into this field. This study attempts to add to the larger conversation about how family backgrounds influence academic outcomes by exploring the complex associations between family background and student academic outcomes.

The first question this study pursues is how the student's academic performance and study habits differ between Mathematics and Portuguese. These are important subjects in Portugal’s educational system, even though they require different kinds of skills. While analytical and logical reasoning and problem solving are the prerequisites for Mathematics, communication, and linguistic internals are the necessities of the Portuguese language. Knowing the special challenges and strengths of students in each subject permits educators to create strategies that take into account each unique learning need.

The second question this study explores is whether family support impacts Math and Portuguese performance differently. It has been well established that family support is

core to academic success, but its role may differ between subjects. Take Mathematics for instance, where structural factors, such as parents' education level or participation in academic activities, could have a stronger impact because Mathematics requires logical and analytical thinking. On the other hand, emotional support, like having supportive family relationships and encouragement, might be more important for learning Portuguese in terms of communication and language expressions used. Understanding how various types of support affect performance in specific subjects can direct parents and instructors on how to best support their students' needs. In turn, this could encourage a more positive environment for learning and benefit students in Mathematics and Portuguese classes.

Examining these relationships can help educators to design more specific strategies that boost student educational outcomes. In the next section, our group wires into the data and performs an analysis, digging to see the patterns and trends that manifest in the dataset.

Data

We obtained the original dataset from Kaggle, which includes two files, `math.csv` and `portuguese.csv`, collected by Paulo Cortez, an Information Systems professor at the University of Minho in Portugal. The data was gathered during the 2005–2006 school year from two public high schools located in the Alentejo region of Portugal. To build the datasets, Cortez combined information from school reports, which relied on paper records, with data collected through questionnaires. The questionnaires were designed with closed-ended questions addressing demographic, social, emotional, and school-related factors thought to influence student performance. After testing the questionnaire on a group of 15 students, it was distributed to 788 students, and the responses were merged with the school records. Following this process, 111 incomplete responses were removed, resulting in two datasets—one for Mathematics (395 observations) and another for Portuguese (649 observations).

Each row in the dataset represents a student from one of two public high schools in the Alentejo region of Portugal during the 2005–2006 school year. The students' data includes academic performance metrics, such as grades across three periods and the number of school absences, alongside demographic, social, and emotional factors collected through questionnaires. The sample consists of 395 students for Mathematics and 649 students for Portuguese, forming two separate datasets. By integrating school reports with the questionnaire responses, the dataset provides a rich snapshot of

student performance, influenced by both academic and non-academic variables.

Before analysis, we carefully examined the dataset for any inconsistencies or missing values across its 32 columns. Remarkably, no missing values were found, indicating that the data was complete. We also implemented a custom method to detect potential outliers in the dataset. However, we decided to retain these outliers since each row represents a unique student, and the dataset is relatively small, with only 395 observations in the math dataset and 649 in the Portuguese dataset. Removing outliers could have led to a loss of valuable information. To prepare for analysis, we merged the two datasets—one containing information on mathematics and the other on Portuguese—because some students were enrolled in both subjects. Students were matched across the datasets using common attributes, such as their school, gender, age, and parental information. During this process, it was observed that 25 students appeared only in the mathematics dataset and did not have corresponding records in the Portuguese dataset. These unmatched students were excluded to ensure consistency in the merged dataset. Duplicates were also removed, leaving a clean, combined dataset where each row represents a unique student with their combined math and Portuguese-related attributes. The final merged dataset retained key variables from both subjects and included unique identifiers for each student to facilitate further analysis. This comprehensive integration provided a complete view of the factors influencing student performance across both core subjects.

The primary target variables in this analysis are the final grades in Mathematics (`G3_math`) and Portuguese (`G3_port`). These were originally numerical variables, ranging from 0 to 20. To simplify the analysis and address classification tasks, we transformed these numerical variables into binary categories: students with final grades of 10 or above were classified as "Pass" (1), while those with grades below 10 were classified as "Fail" (0). This transformation aligns with common educational benchmarks, making it easier to interpret whether a student meets the minimum academic success threshold. To address class imbalance in the binary target variables, we applied SMOTE (Synthetic Minority Oversampling Technique), a popular oversampling method. SMOTE works by generating synthetic samples for the minority class (in this case, "Fail"), rather than duplicating existing samples. It creates new data points by interpolating between existing minority class samples, ensuring that the distribution of the minority class is more representative and avoids overfitting. By balancing the target variable classes, SMOTE helps improve the performance and fairness of classification models, enabling them to learn from both classes more effectively.

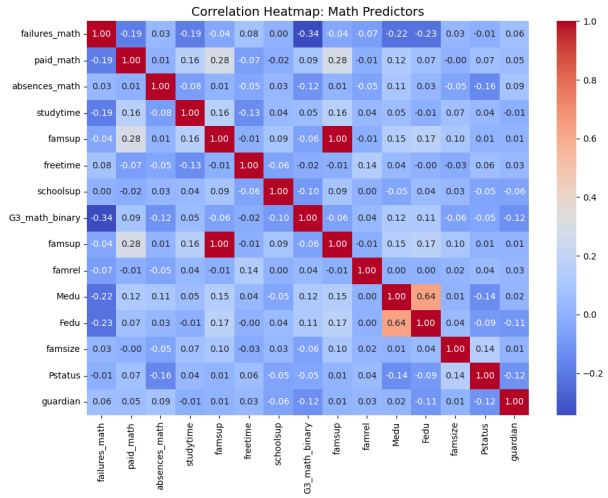
Class Distribution of Target Variables Before SMOTE

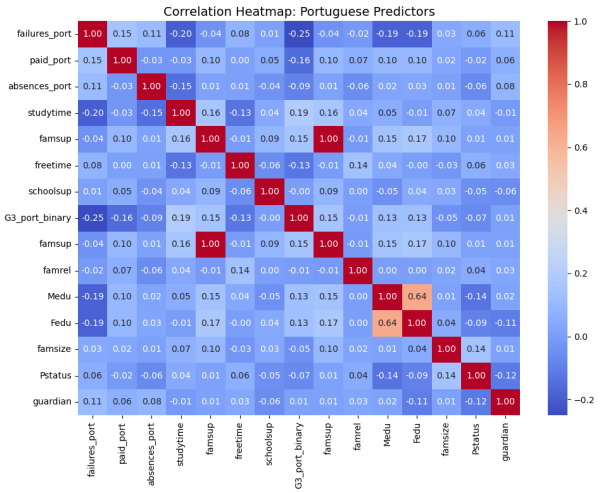
Class	Math	Portuguese
Pass (1)	249	341
Fail (0)	121	29

Class Distribution of Target Variables After SMOTE

Class	Math	Portuguese
Pass (1)	249	341
Fail (0)	249	341

The predictor variables used in this study vary depending on the research questions. For Question 1, which examines subject-specific factors influencing success in Mathematics and Portuguese, we selected predictors such as `failures` (past class failures), `paid` (participation in extra paid classes), `absences`, `studytime`, `famsup` (family support), `freetime`, and `schoolsup` (educational support). These variables were subject-specific, with separate versions for Mathematics (e.g., `failures_math`) and Portuguese (e.g., `failures_port`). For Question 2, focusing on broader familial and social influences, the predictors included `famsup` (family support), `famrel` (family relationship quality), `Medu` and `Fedu` (mother's and father's education levels), `famsize` (family size), `Pstatus` (parental cohabitation status), and `guardian` (primary caregiver). To simplify modeling and ensure consistency, categorical predictor variables, such as `famsup`, `Pstatus`, and `famsize`, were converted into binary representations (e.g., "yes" as 1 and "no" as 0), enabling effective use in classification models. To further refine the model-building process, correlation heatmaps were generated to examine the relationships between predictors and the target variables (`G3_math_binary` and `G3_port_binary`). These heatmaps reveal several key patterns. For example, `failures_math` and `failures_port` are strongly negatively correlated with the target variables, indicating that past academic failures are significant predictors of student performance. Conversely, variables like `paid_math` and `paid_port` exhibit weak positive correlations, suggesting that extra paid classes provide a modest benefit. The heatmaps also highlight relationships among predictors themselves, such as a strong correlation between `Medu` and `Fedu`, which may indicate redundancy. These insights help guide feature selection by prioritizing relevant variables, identifying potential multicollinearity, and improving the interpretability and stability of the models. As a result, the heatmaps serve as a crucial exploratory tool in the data preprocessing and modeling pipeline.





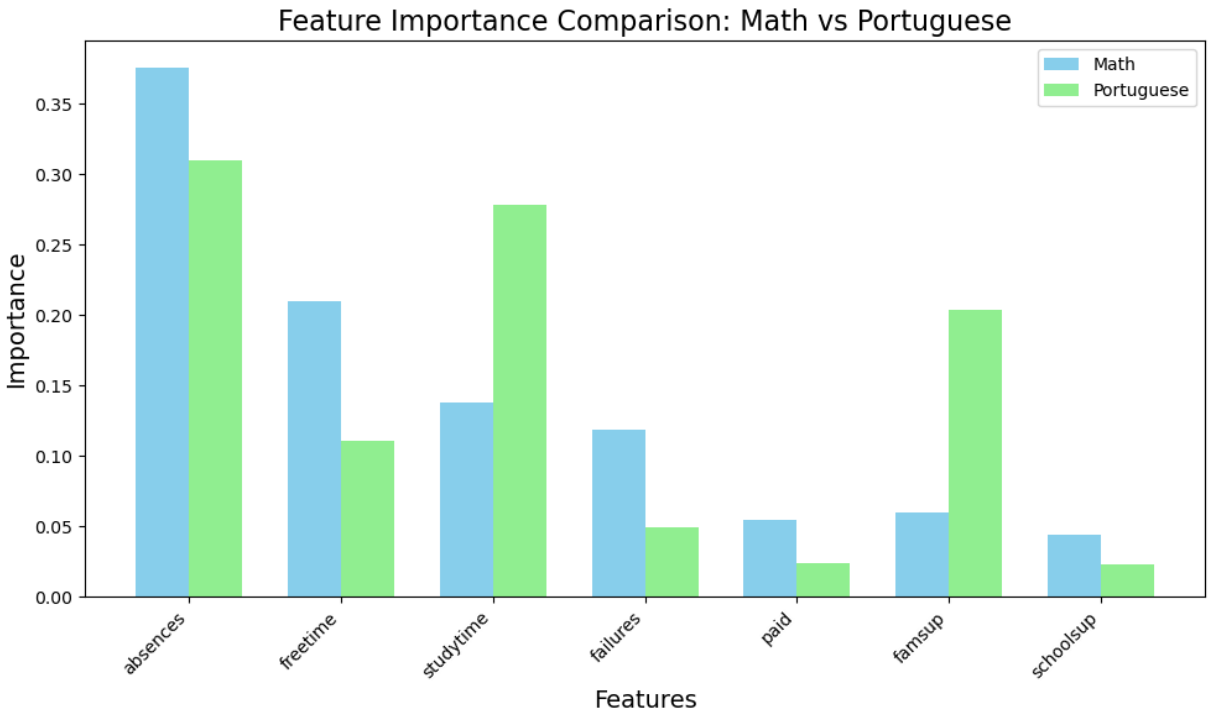
Results

Problem 1

To address the first research question, a Random Forest Classifier was chosen for its ability to handle complex relationships among predictors and target variables. The predictor variables included features related to students' study habits and support systems, such as `failures_math`, `paid_math`, `absences_math`, `studytime`, `famsup`, `freetime`, and `schoolsup` for Mathematics, and their respective counterparts for Portuguese. The target variables were the binary indicators for passing or failing (`G3_math_binary` and `G3_port_binary`), derived from students' final grades. The data was split into training and testing sets after applying SMOTE to balance the class distribution, addressing the initial imbalance where the "fail" class was underrepresented.

Class	Precision (Math)	Recall (Math)	F1-Score (Math)	Support (Math)	Precision (Portuguese)	Recall (Portuguese)	F1-Scor (Portugue
0	0.68	0.72	0.70	50	0.65	0.83	0.73
1	0.70	0.66	0.68	50	0.87	0.73	0.79
Accuracy	**0.69**	**-**	**-**	**100**	**0.77**	**-**	**-**
Macro Avg	0.69	0.69	0.69	100	0.76	0.78	0.76
Weighted Avg	0.69	0.69	0.69	100	0.79	0.77	0.77

Based on the table, the Random Forest model for Mathematics achieved an overall accuracy of 69%, with a precision of 68% for the "fail" class and 70% for the "pass" class. The recall scores, indicating the model's ability to identify true positives, were 72% for the "fail" class and 66% for the "pass" class. This shows that the model is slightly better at identifying students likely to fail but struggles with false positives. For Portuguese, the Random Forest model performed better, with an overall accuracy of 77%. The precision for the "fail" class was 65%, with a recall of 83%, indicating the model effectively identified most failing students. The precision and recall for the "pass" class were 87% and 73%, respectively. The higher recall for the "fail" class in Portuguese suggests that the model was more sensitive to underperforming students in that subject. Overall, the model tends to predict slightly more students to fail who actually pass, the Portuguese model is better at predicting passing students, and the models are decent at balancing precision and recall (F1-score) for predicting student outcomes. This is helpful to our analysis because we know that misclassification risks and class imbalances are no longer major issues in our data.



The visualized feature importance through bar plots highlights distinct differences between the predictors for Mathematics and Portuguese performance. For Mathematics, absences emerge as the most influential feature, underscoring the substantial impact of attendance on academic outcomes in this subject. This emphasizes that consistent classroom engagement is critical for mastering mathematical concepts, which often require step-by-step learning and regular practice. In contrast, for Portuguese, the

feature importance is more balanced, with absences, study time, and family support (`famsup`) contributing significantly. This suggests that Portuguese performance may be influenced by a broader combination of factors, including family involvement, personal time management, and attendance. The higher importance of study time for Portuguese reflects the subject's emphasis on self-directed learning, reading, and writing, which require extended periods of focused effort outside the classroom. The relatively higher importance of `famsup` for Portuguese compared to Mathematics further implies that family environment plays a more critical role in language subjects. This may stem from the need for consistent encouragement, resources, and support at home to develop language skills. In comparison, Mathematics appears to be driven more by individual effort and attendance, as seen from the higher influence of absences and the lesser role of outside support.

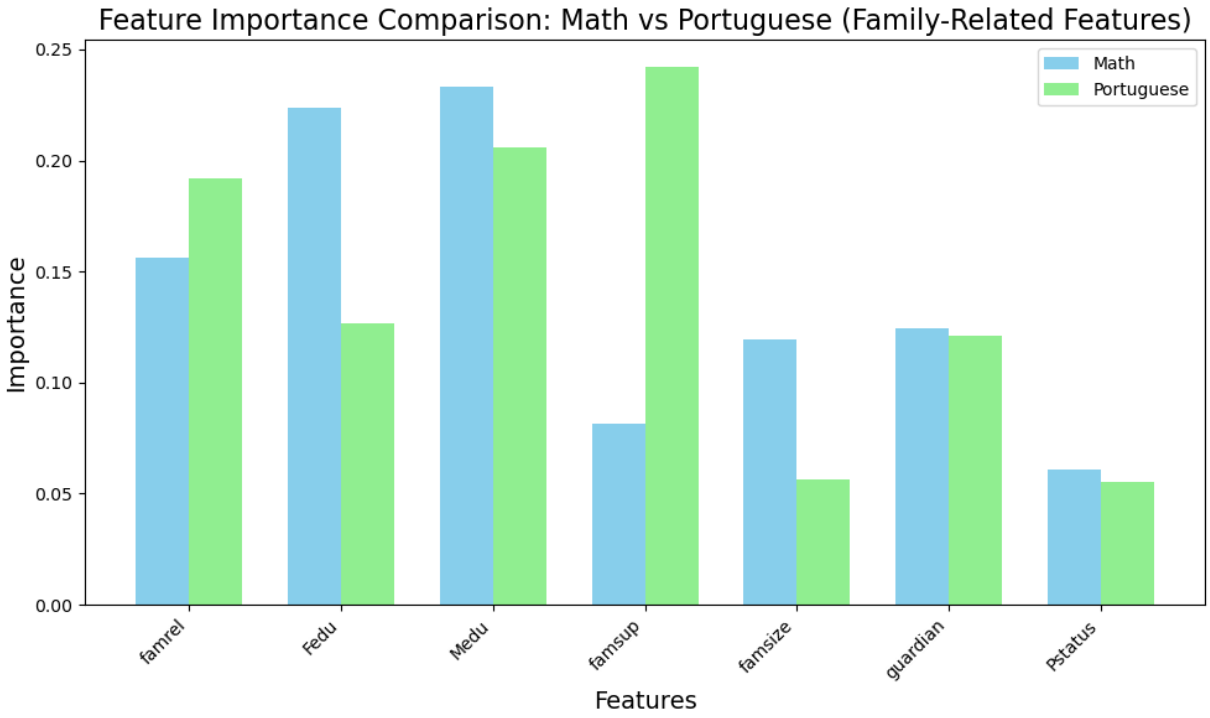
Additionally, other factors such as failures in Mathematics and free time show differences in their contributions. Failures in Mathematics are an important indicator, suggesting that overcoming past academic challenges in this subject significantly impacts performance. Conversely, free time plays a relatively balanced role in Portuguese, potentially pointing to the benefits of relaxation and extracurricular activities for overall well-being and academic performance. Overall, the visualization demonstrates that Mathematics and Portuguese require distinct approaches to optimize performance. While Mathematics relies heavily on consistent effort and attendance, Portuguese benefits from a more holistic approach involving family support, study time, and a balanced personal life.

The findings also reveal actionable insights for educators and policymakers. For Mathematics, interventions targeting absenteeism, such as real-time attendance monitoring and follow-up support for missed lessons, could be effective in improving student outcomes. Additionally, promoting effective time management strategies and providing additional resources for self-study may also enhance performance. For Portuguese, fostering parental engagement programs and workshops to involve families more deeply in students' education could be a valuable strategy. The identification of these subject-specific influences enables schools to design tailored interventions that cater to the unique challenges of each subject, thereby enhancing the overall academic success of students.

Problem 2

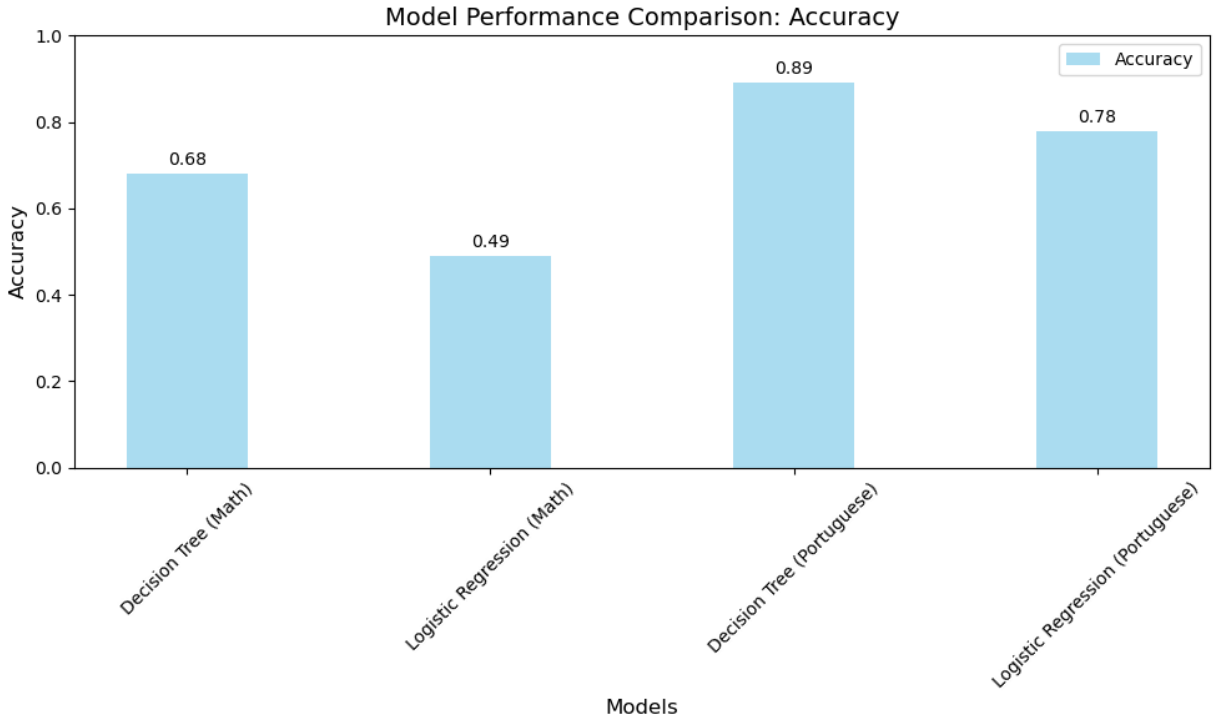
The objective of this analysis is to investigate whether family support correlates differently with student performance in Math versus Portuguese, and to identify the

specific aspects of family support that have the greatest impact. To address this question, two models—a Decision Tree Classifier and Logistic Regression—were employed. These models allow for an exploration of both feature importance and the predictive performance of family-related variables, such as parental education, family size, and relationship quality, on the binary classification outcomes of "pass" and "fail" for each subject.



Using the Decision Tree Classifier, the family-related predictor variables (e.g., parental education levels `Medu` and `Fedu`, family relationship `famrel`, and family support `famsup`) were analyzed for their contributions to predicting Math and Portuguese performance. For the Math dataset, the optimal tree depth was determined to be 12, based on cross-validation scores. The test accuracy at this depth was 68%, and the feature importance analysis revealed that `Medu` (mother's education) and `Fedu` (father's education) were the most influential predictors, followed by `famrel`. In the case of Portuguese, the same Decision Tree approach was applied, resulting in an optimal depth of 12 with a test accuracy of 89%. Interestingly, the importance of `famsup` (family educational support) emerged as the highest, followed by `Medu` and `famrel`. This indicates a stronger association between direct family support and performance in Portuguese compared to Math. The higher accuracy achieved for Portuguese suggests that family-related variables may have a more substantial and direct impact on Portuguese performance.

The Logistic Regression model was also applied to analyze the influence of family-related predictors. For Math, regularization parameter tuning identified the best value of C as 0.001, optimizing the balance between underfitting and overfitting. However, this may have led to oversimplifying the model, leading to poor performance. The model achieved a test accuracy of 49%, with relatively low precision and recall scores, indicating limited predictive capability for this dataset, performing no better than random guessing. The ROC curve for Math showed an Area Under the Curve (AUC) of 0.49, further suggesting poor performance. Despite this, the classification report indicated that the model was slightly better at predicting "fail" outcomes compared to "pass." For Portuguese, the Logistic Regression model performed significantly better, achieving a test accuracy of 78% with the best C value determined to be 46.416, allowing the model to capture more nuanced patterns. The precision, recall, and F1 scores were markedly higher compared to Math, with an AUC of 0.89, highlighting strong discriminatory ability. The classification report indicated that the model was particularly effective at identifying "pass" outcomes. The findings emphasize that Logistic Regression, while less interpretable than the Decision Tree, can still offer valuable insights into the linear relationships between family-related variables and Portuguese performance.



The Decision Tree model provided interpretable insights into the importance of individual features, highlighting "Medu" and "Fedu" for Math and "famsup" for Portuguese as the most significant predictors. In contrast, Logistic Regression revealed that linear relationships between variables were more effective for Portuguese, as evidenced by the

higher test accuracy and AUC. However, the Logistic Regression model struggled with the Math dataset, possibly due to nonlinear relationships or insufficient feature representation. Overall, the Decision Tree model outperformed Logistic Regression for both Math and Portuguese. Specifically, it achieved a test accuracy of 68% for Math and 89% for Portuguese, compared to Logistic Regression's 49% and 78%, respectively. This demonstrates that the Decision Tree was more effective in capturing the underlying patterns in the data for both subjects.

This analysis demonstrates that family support factors, such as parental education and direct family educational support, influence student performance in Math and Portuguese differently. For Math, parental education appears to be the most critical factor, while for Portuguese, direct family support plays a more significant role. These findings directly answer Question 2 by identifying the specific aspects of family support that correlate differently with performance in these subjects. Targeted interventions focusing on improving family support systems, such as parent education programs and direct family engagement in students' learning processes, may enhance academic outcomes, particularly in Portuguese. Future work could explore more advanced models or additional features to further refine these findings.

Conclusion

In this study, we explored how various factors influence student performance in Mathematics and Portuguese, two core subjects in Portugal. Our analysis of data from two public high schools aimed to answer how 1) academic performance and study habits differ between these subjects, and 2) how family support impacts each. We have found that attendance, study time, parental education levels, and the number of past failures significantly affect outcomes. In Mathematics, structural factors such as parental education, study habits, and self control played a stronger role, while in Portuguese, emotional support from family was more influential. For example, students whose parents had higher education levels excelled in Mathematics, whereas positive family relationships correlated with better Portuguese performance. Our findings were expected, but also presented some nuances. Specifically, the extent to which absences dominated the predictors was surprising, and parental education levels being more critical for Math and familial support for Portuguese highlight how even family factors have their academic and emotional facets.

These findings are important because they provide actionable insights for educators, parents, and policymakers. Recognizing that different subjects require distinct types of

support enables the creation of tailored educational strategies. Schools can design programs such as logical reasoning workshops for Mathematics and family communication-focused activities for Portuguese. Parents can better understand how their involvement, whether through academic support or emotional encouragement, directly influences their child's success. Policymakers can allocate resources to support both academic interventions and family engagement initiatives, ensuring a comprehensive approach to improving educational outcomes. These results can be applied in real-world educational settings to address specific challenges. For example, targeted interventions for students with repeated failures can help them get back on track. Parental education programs can be developed to highlight the importance of both structural and emotional support in different subjects. Additionally, teacher training programs can emphasize recognizing subject-specific challenges and adopting effective teaching methods. By implementing these strategies, schools can create a more supportive environment that enhances student achievements.

This study provides valuable insights, but there is ample room for further exploration. We later discovered a moderate correlation between Math and Portuguese performance (0.49), which might suggest shared underlying factors and potential multicollinearity we didn't account for. This may have increased the risk of overfitting in our models, particularly when interpreting feature importances or fixing class imbalances. Future models can take this into account and explicitly model their interdependence, which might lead to more distinct predictions and interpretations of subject-specific influences. Predicting final (third quarter) grades with first and second quarter grades is also an option that may increase our model accuracies. Future research can improve the class imbalance issue by gathering more data or exploring alternative methods, such as ensemble techniques or cost-sensitive learning, to complement SMOTE. Our Random Forest and Logistic Regression models can also definitely be improved using hyperparameter tuning or algorithms like Gradient Boosting or Neural Network models. Other methods like factor analysis or principal component analysis might discover latent structures in the data, which could improve interpretability as well. Including additional data sources, such as socio-economic variables, mental health indicators, peer influence metrics, and teacher effectiveness could provide deeper insights. Collecting longitudinal data would also help analyze performance trends over time. These improvements could refine our understanding of students' success, paving the way for more effective, data-driven educational policies.