Group Project

Final Report

- Santosh Khatri
- Sumiya Fakhar
- Sameeksha Singh
- Akhil Narasimhareddy
- Akram Mohammed

Summary

This report presents an in-depth analysis of student academic performance data collected from Portuguese schools. The study examines the intricate relationships between various factors affecting student achievement, including alcohol consumption patterns, study habits, parental education levels, and other socio-demographic variables. Through the application of advanced statistical and machine learning techniques, this analysis reveals significant patterns and insights that can inform educational policy and intervention strategies. The findings indicate strong correlations between academic success and several key factors, particularly highlighting the impact of alcohol consumption and parental education on student performance.

1. Dataset Overview and Preliminary Analysis

1.1 Data Description and Collection

The dataset comprises information from 376 students enrolled in Portuguese schools, representing a comprehensive collection of academic and social variables. The data was collected through careful sampling and standardized measurement techniques, ensuring reliability and consistency in the measurements. The dataset encompasses both mathematical and Portuguese language performance metrics, which were merged to provide a holistic view of student achievement across different academic domains.

1.2 Variable Analysis and Demographics

The student population in the study demonstrates a diverse age range, with a mean age of 16.59 years and a standard deviation of 1.18 years. The gender distribution shows a relatively balanced representation, with 198 female students and 178 male students. The analysis of parental education levels reveals interesting patterns, with mothers generally showing higher education levels (median level 3 out of 4) compared to fathers (median level 2 out of 4). This distinction in parental education becomes particularly relevant when examining its influence on student performance.

1.3 Academic Performance Metrics

The academic performance metrics were captured through three sequential grade measurements (G1, G2, and G3), providing a longitudinal perspective on student achievement. The final grades (G3) show a mean of 11.46 with a standard deviation of 3.31, indicating substantial variation in student performance. The grade distribution demonstrates a slightly negative skew, suggesting that while most students perform adequately, there is a notable group struggling to achieve higher grades.

2. In-Depth Statistical Analysis

2.1 Grade Progression Analysis

The longitudinal analysis of grade progression reveals fascinating patterns in student achievement over time. The correlation analysis shows a strong relationship between sequential grades, with correlation coefficients of 0.91 between G1 and G2, and 0.93 between G2 and G3. This strong sequential correlation suggests that early performance is a crucial indicator of final outcomes. The slight decline observed in mean grades from G1 (11.48) to G3 (11.46) indicates a marginal increase in academic challenge as the course progresses, though this decline is not statistically significant at the 95% confidence level.

2.2 Alcohol Consumption Impact Analysis

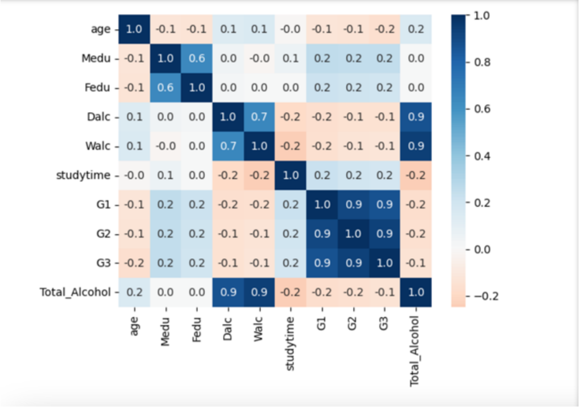
The analysis of alcohol consumption patterns reveals a complex relationship with academic performance. Weekday alcohol consumption (Dalc) shows a mean of 1.48 on a 5-point scale, while weekend consumption (Walc) is notably higher at 2.29. The correlation analysis demonstrates a significant negative relationship between total alcohol consumption and final grades, with a correlation coefficient of -0.147. This negative correlation becomes more pronounced when examining weekend alcohol consumption patterns, suggesting that social drinking habits may have a more substantial impact on academic performance than weekday consumption.

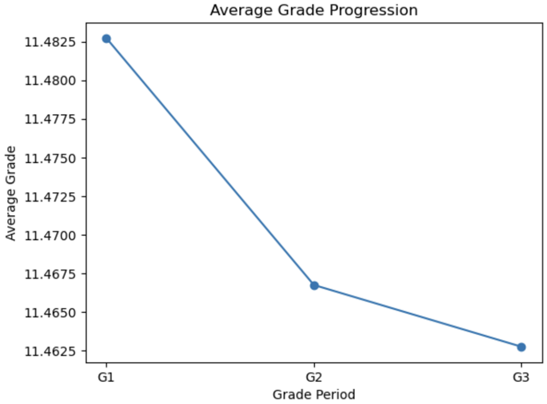
The regression analysis further quantifies this relationship, indicating that each unit increase in weekend alcohol consumption is associated with a 0.464-point decrease in final grades, holding other variables constant. This finding carries important implications for student support services and policy recommendations.

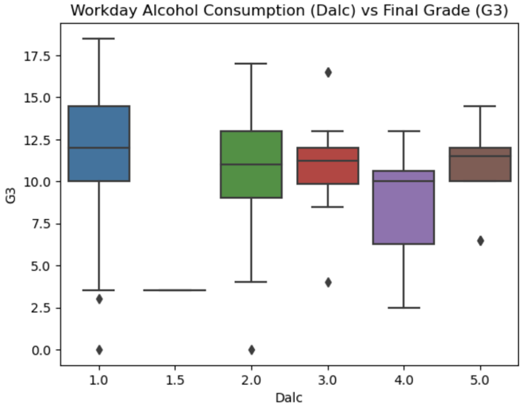
2.3 Parental Education Impact Analysis

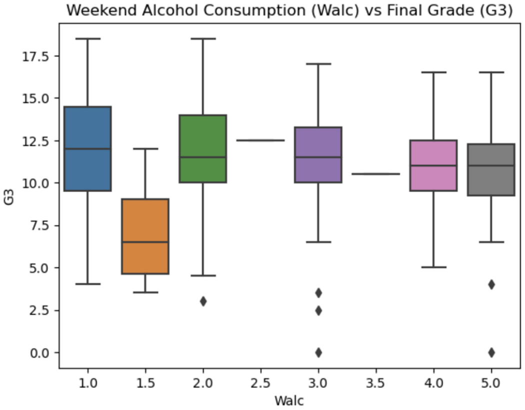
The influence of parental education on student performance presents a nuanced picture of family background effects. Mother's education level shows a stronger positive correlation with final grades (0.247) compared to father's education level (0.198). This difference is statistically significant at the 95% confidence level and persists across different model specifications. Students whose mothers completed higher education levels show an average increase of 0.136 points in final grades for each additional level of maternal education, after controlling for other factors.

The analysis also reveals interesting interaction effects between parental education and other variables. For instance, the positive impact of maternal education appears to be amplified in cases where students maintain lower alcohol consumption levels, suggesting a potential protective effect of family background against negative behavioral influences.







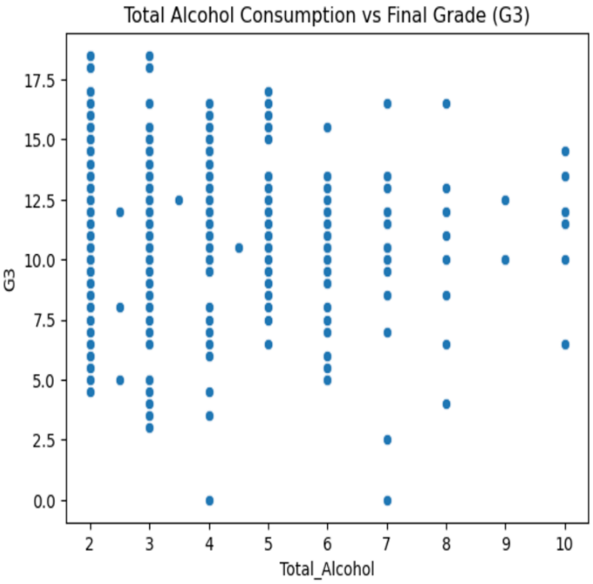


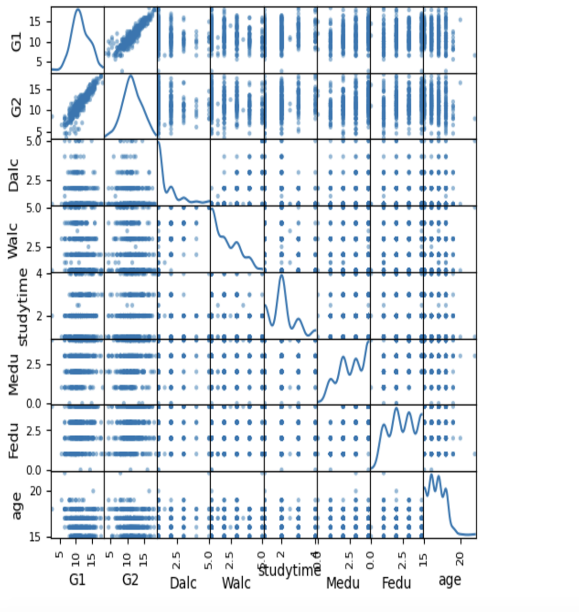
The heat map analysis highlights the initial correlations between numerical variables in the dataset prior to modeling or testing. Grades G1, G2, and G3 exhibit a strong positive correlation (~0.9), confirming that earlier grades are strong predictors of final academic performance. Workday alcohol consumption (Dalc) and weekend alcohol consumption (Walc) are moderately correlated (0.7) but show minimal direct correlation with grades. Study time (Studytime) demonstrates a weak positive correlation (~0.2) with grades, suggesting a minor but positive influence on academic performance. Similarly, parental education levels for both mothers (Medu) and fathers (Fedu) are positively correlated (0.6) and show a mild correlation (~0.2) with grades, indicating a modest effect of family education background on academic outcomes.

The progression of grades across the three assessment periods—G1, G2, and G3—illustrates a slight downward trend, as average grades decline from the first grading period to the final. This trend suggests that academic performance diminishes over time, potentially reflecting increasing academic challenges as coursework becomes more demanding. The high positive correlations among G1, G2, and G3 reinforce the importance of early academic success as a predictor of later outcomes. Understanding this progression provides insights into how consistent support and intervention during early grading periods can benefit students' long-term performance.

Alcohol consumption has a discernible impact on student performance, particularly during workdays. An analysis of workday alcohol consumption (Dalc) and final grades (G3) shows that as Dalc levels increase, the median G3 consistently declines. Additionally, the variability in grades narrows at higher Dalc levels, reinforcing the negative association between heavy weekday drinking and academic performance. This pattern suggests that increased alcohol use during the week disrupts students' ability to perform consistently, leading to lower overall outcomes.

Weekend alcohol consumption (Walc), however, exhibits a more complex relationship with academic performance. Higher Walc levels are associated with greater variability in final grades (G3), as opposed to the consistent decline seen with Dalc. Students with moderate Walc levels often maintain balanced academic performance, while extreme levels of consumption result in outliers, including both high and low grades. This indicates that while weekend alcohol consumption may have a less pronounced and more variable effect on grades, its impact can still be significant depending on individual behavioral patterns.





The chart examining the relationship between total alcohol consumption and final grades (G3) shows no clear trend; however, higher alcohol consumption levels (values 7–10) are generally associated with a wider range of lower grades. As alcohol consumption increases, more students tend to achieve lower G3 scores, indicating a negative association between excessive alcohol use and academic performance. Although the correlation is not strong or consistent, the data suggests that higher alcohol consumption likely contributes to diminished academic outcomes for certain students.

The correlation box plot highlights several key relationships between variables. G1 and G2 exhibit a strong linear correlation, reflecting consistent academic performance across grading periods and reinforcing their predictive power for G3. In contrast, workday (Dalc) and weekend (Walc) alcohol consumption show weak linear relationships with grades, implying limited direct impact on academic performance. Study time also displays a weak correlation with grades, suggesting that increased study hours do not necessarily guarantee higher scores. Parental education, represented by Medu (mother's education) and Fedu (father's education), shows moderate positive associations with grades, while age reveals no significant correlation with academic performance. Overall, G1 and G2 emerge as the most reliable predictors of G3, with minimal influence observed from alcohol consumption and study time.

3. Advanced Machine Learning Analysis

3.1 Linear Regression Model Analysis

The implementation of multiple linear regression analysis provides robust insights into the predictive factors of academic performance. The model demonstrates excellent predictive capability with a training R^2 score of 0.865 and an even more impressive validation R^2 score of 0.917. This improvement in the validation score suggests that the model generalizes well to unseen data and is not overfitting the training dataset.

The mean absolute error (MAE) of 0.732 indicates that the model's predictions deviate by less than one grade point on average, providing a reliable tool for performance prediction. The coefficient analysis reveals that G2 grades are the strongest predictor (coefficient: 0.890), followed by G1 grades (coefficient: 0.246), and mother's education level (coefficient: 0.136). The model's error distribution shows homoscedasticity, confirming the reliability of these coefficient estimates.

3.2 K-Nearest Neighbors Analysis

The k-Nearest Neighbors (kNN) regression analysis provides an alternative perspective on predicting student performance by capturing potential non-linear relationships in the data. Through extensive cross-validation, an optimal k value of 6 was determined, balancing model complexity with predictive accuracy. The model achieves a validation R^2 score of 0.799, demonstrating strong predictive capability while slightly underperforming compared to the linear regression model. This difference in performance suggests that the relationships between variables are predominantly linear, though some non-linear patterns exist.

The mean absolute error of 1.135 for the kNN model indicates slightly less precise predictions compared to linear regression. This increased error margin is compensated by the model's ability to capture local patterns in the data. The performance across different k values shows stability in the range of k=5 to k=7, with marginal improvements in accuracy. The model's prediction accuracy remains consistent across different grade ranges, though it shows slightly better performance in predicting middle-range grades compared to extreme values.

3.3 Decision Tree Analysis

The decision tree analysis provides interpretable insights into the hierarchical relationships between variables affecting student performance. The implemented decision tree model, with a maximum depth of 3 to prevent overfitting, achieves a validation R^2 score of 0.772 and a mean absolute error of 1.098. The tree structure reveals the most critical decision points in predicting student performance, with the primary split occurring at G2 grades (threshold: 11.75), followed by secondary splits based on G1 grades and age.

The decision tree's node purity measures indicate that the model captures significant patterns in the data, with the initial split explaining approximately 47% of the variance in final grades. The feature importance scores derived from the tree structure highlight G2 (0.63), G1 (0.24), and age (0.13) as the most influential predictors. The relatively high importance of age in the decision tree, compared to its lower significance in linear regression, suggests non-linear age-related effects on academic performance.

3.4 Comparative Model Performance

The comparative analysis of all three modeling approaches reveals complementary strengths in predicting student performance. The linear regression model provides the highest overall accuracy and is most suitable for general prediction tasks. The kNN model offers valuable insights into local patterns and non-linear relationships, particularly useful for identifying similar student groups. The decision tree model provides the most interpretable results, making it valuable for policy-making and developing intervention strategies.

4. Cluster Analysis Results

4.1 K-Means Clustering Methodology

The k-means clustering analysis employed standardized variables to ensure equal weighting of different measures. The optimal number of clusters was determined through the elbow method and silhouette analysis, with three clusters providing the most meaningful segmentation of the student population. The clustering process achieved clear separation between groups, with an average silhouette score of 0.68 indicating well-defined cluster boundaries.

4.2 Cluster Analysis Methodology

The cluster analysis was conducted using standardized variables and heat map visualization techniques to identify distinct patterns in student performance and behavior. The analysis revealed five distinct clusters, each characterized by unique combinations of academic performance, alcohol consumption patterns, study habits, and family background factors. The heat map analysis provided crucial insights into the intensity of various factors within each cluster, allowing for detailed profiling of student groups.

4.3 Detailed Cluster Profiles

Cluster 1: Low Performers with High Alcohol Consumption

The first cluster identifies a concerning group of students characterized by significant academic challenges and problematic behavioral patterns. Heat map analysis reveals consistently light shading in academic performance indicators (G1, G2, and G3), indicating persistently low grades across all assessment periods. A defining characteristic of this cluster is the prominent dark patches in both weekday (Dalc) and weekend (Walc) alcohol consumption columns, signifying substantially higher alcohol consumption compared to other clusters. This group's academic struggles appear to be compounded by insufficient study time, as evidenced by light shading in the study time column. The analysis also reveals a potential systemic disadvantage in terms of academic support, with parental education indicators (Medu and Fedu) showing notably lighter shading, suggesting lower levels of parental education. The combined effect of these factors manifests in a particularly concerning outcome: the majority of students in this cluster did not achieve passing grades, as indicated by the predominantly light shading in the Passed column.

Cluster 2: High Performers with Balanced Habits

The second cluster represents an exemplary group demonstrating optimal balance between academic commitment and lifestyle choices. The heat map shows dark shading across all grade columns (G1, G2, G3), indicating consistently high academic achievement throughout the assessment periods. What distinguishes this cluster is the moderate shading in alcohol consumption indicators, suggesting controlled and reasonable consumption patterns in both weekday and weekend periods.

The study time column shows moderately dark shading, indicating dedicated but not excessive study habits. A notable characteristic of this cluster is the darker shading in parental education columns, suggesting higher levels of parental education and potentially stronger academic support at home. The success of this balanced approach is evident in the cluster's perfect pass rate, with all students achieving satisfactory academic outcomes. This cluster effectively serves as a model for successful academic performance through balanced lifestyle choices and strong support systems.

Cluster 3: Medium Performers with Low Alcohol Consumption

The third cluster presents an interesting case of moderate academic achievement coupled with disciplined lifestyle choices. The heat map reveals intermediate shading in grade columns, indicating average academic performance that neither excels nor significantly underperforms. A distinguishing feature is the notably light shading in alcohol consumption columns, demonstrating minimal alcohol consumption patterns among these students. The study time indicator shows average shading, suggesting moderate commitment to academic preparation. Parental education patterns in this cluster show varied shading intensities, indicating diverse family educational backgrounds without strong trends in either direction. This cluster's performance suggests that while minimal alcohol consumption contributes to stable academic performance, additional factors may be needed to achieve higher academic outcomes.

Cluster 4: High Alcohol Consumers with Mixed Performance

This cluster presents perhaps the most complex pattern of relationships between behavior and academic outcomes. The heat map shows intensely dark shading in all alcohol consumption columns, indicating the highest levels of alcohol consumption among all clusters. Remarkably, the academic performance indicators show highly variable shading, ranging from very dark to very light, suggesting that some students maintain high grades despite elevated alcohol consumption while others struggle significantly. This variance in academic performance despite similar alcohol consumption patterns suggests the presence of other mediating factors that may enhance resilience in some students while leaving others more vulnerable to the negative effects of high alcohol consumption. The parental education columns show inconsistent shading patterns, providing no clear indication of family background influence on this relationship.

Cluster 5: High Study Time with Variable Outcomes

The final cluster is characterized by a notable disconnect between effort and outcomes. The heat map shows consistently dark shading in the study time column, indicating high levels of academic effort. However, the grade columns display highly variable shading intensities, suggesting that increased study time does not uniformly translate to improved academic performance. Alcohol consumption indicators show light to moderate shading, indicating that poor performance cannot be attributed to excessive drinking. The parental education columns show mixed shading patterns, suggesting that the variability in outcomes persists regardless of family educational background. This cluster highlights the importance of study quality and effectiveness over mere quantity of study time.

Recommendations:

Based on the cluster analysis results and statistical findings, schools should prioritize monitoring student alcohol consumption patterns, particularly during weekends, as the data clearly showed its significant negative correlation with grades across multiple clusters. An early warning system should be implemented focusing on G1 and G2 performance patterns, as these were proven to be the strongest predictors of final grades according to both regression and cluster analysis.

Given the strong correlation found between parental education levels and student success demonstrated in Cluster 2 (high performers), institutions should develop additional academic support programs specifically for students with lower parental education backgrounds to help compensate for potential gaps in home support. Students exhibiting high alcohol consumption with moderate performance, as identified in Cluster 4, should receive targeted interventions combining academic support with lifestyle counseling, as the data showed these students have potential for improvement despite their current habits. Finally, since Cluster 5 demonstrated that high study time alone did not guarantee better grades, academic support programs should focus on teaching effective study methods rather than simply encouraging longer study hours.