

Alcohol Consumption and Student Achievement: A Multi-Model Regression Comparison of Predictors of Final Grades

Anushka Basu Sherelle Li Yifeng Jin

2025-12-17

Contents

1	Introduction & Executive Summary	3
1.1	Executive Summary	3
2	Data Description	3
2.1	Dataset Selection and Source	3
2.2	Outcome Variable	3
2.3	Predictor Domains	4
2.4	Modeling Readiness	4
3	Data Cleaning and Preparation	4
3.1	Data Loading and Outcome Definition	4
3.2	Variable Typing and Encoding Decisions	4
3.3	Feature Engineering: Total Alcohol Consumption	4
3.4	Distribution Checks and Outlier Assessment	5
4	Exploratory Data Analysis (EDA)	5
4.1	Outcome Distribution	5
4.2	Academic and Family Factors	6
4.3	Social and Lifestyle Variables	6
4.4	Alcohol Consumption Patterns	6
4.5	Multicollinearity Assessment	6
4.6	EDA Takeaways	7
5	Methodology	7
5.1	Principal Components Analysis (PCA)	7
5.1.1	PCA Inputs and Preprocessing	7
5.1.2	PCA Estimation	7
5.2	Linear, Stepwise, and Penalized Regression	8
5.2.1	Baseline Linear Regression	8
5.2.2	Stepwise Regression (AIC and BIC)	8
5.2.3	Penalized Regression (Ridge, LASSO, Elastic Net)	9
5.3	Random Forest	9
5.4	XGBoost	9
5.5	Neural Networks & Text Mining	10
6	Results	10
6.1	Principal Components Analysis	10
6.1.1	Implications for Modeling	10
6.2	Linear, Stepwise, and Penalized Regression	10
6.2.1	Model Evaluation and Comparison	11
6.2.2	Implications for Alcohol Consumption (Talc)	12
6.3	Random Forest	13

6.3.1	Predicting Student Letter Grade	13
6.3.2	Predicting Student Performance	14
6.4	XGBoost	14
6.4.1	Predicting Student Letter Grade	14
6.4.2	Predicting Student Performance	15
6.5	Neural Networks & Text Mining Results	15
6.5.1	Qualitative Insights from Text Mining	15
6.5.2	Classification Performance (Neural Network vs Tree)	16
6.5.3	Training Dynamics	16
6.5.4	Feature Importance (Bagged Tree)	16
7	Conclusion	16
7.1	Linear, Stepwise, and Penalized Regression	16
7.2	Tree Based Models	17
7.3	Neural Networks & Text Mining	17
7.4	Limitations and Future Directions	17
8	Appendix	18
8.1	Summary Statistics Tables and Selected Predictors Tables	18
8.2	Explanation of XGBoost Algorithm	19

1 Introduction & Executive Summary

Alcohol consumption is a common feature of adolescent and young adult social life, and a substantial body of research has examined its relationship with academic outcomes. Prior studies consistently find that higher levels of alcohol use are associated with lower grades, increased absenteeism, and reduced academic engagement, though the magnitude of these effects varies once background and behavioral factors are taken into account. For example, DeSimone (2010) finds that binge drinking is linked to lower GPA after controlling for student characteristics, while Wolaver (2002) shows that alcohol use is strongly related to missed classes, which in turn affects academic performance (DeSimone, 2010; Wolaver, 2002).

At the same time, alcohol consumption often co-occurs with other factors—such as prior academic performance, study habits, family background, and peer socialization—that are themselves strong predictors of grades. This raises an important question: does alcohol consumption have an independent relationship with academic performance, or does its apparent effect largely reflect underlying differences in behavior and context?

This question is also personally relevant. Drinking as part of party and social culture is prevalent on many college campuses, including our own university, making it especially interesting to examine whether patterns observed in prior research are already evident earlier in students' academic trajectories. Motivated by both the literature and lived experience, this project asks: **How does student alcohol consumption relate to academic performance, and does it meaningfully contribute to predicting final grades beyond prior academic performance and other behavioral and family factors?**

1.1 Executive Summary

This analysis examines how alcohol consumption relates to academic performance among secondary school students, using a rich set of behavioral, social, demographic, and academic variables. After extensive cleaning, exploratory analysis, and feature engineering, multiple regression frameworks were applied—including baseline OLS, stepwise selection (AIC/BIC), and penalized models (Ridge, LASSO, Elastic Net)—to compare predictive accuracy, interpretability, and model stability.

Across all models, academic history (especially prior failures), study behavior, and educational aspirations consistently emerged as the strongest predictors of final grades. Alcohol consumption, summarized through a combined Talc variable, showed a small negative association with performance, but its effect size was modest relative to core academic features. Importantly, penalized regression confirmed that alcohol contributes some predictive signal, while stepwise models demonstrated that it does not substantially improve model fit once dominant variables are included.

Overall, results show that alcohol matters—but not nearly as much as academic effort, engagement, and prior achievement. Final grades are best explained by structural and behavioral factors rather than alcohol use alone, suggesting that alcohol should be interpreted as one component within a broader academic risk profile rather than a primary driver of academic outcomes.

2 Data Description

2.1 Dataset Selection and Source

This study uses the student performance dataset for the Portuguese language course (student-por.csv) from the UCI Machine Learning Repository. The Portuguese dataset was selected instead of the Mathematics course dataset because it contains a larger number of observations ($n = 649$ versus $n = 395$). The larger sample size improves statistical power and supports more stable estimation in multivariate regression and regularization-based models that rely on sufficient data density.

Although the Mathematics and Portuguese datasets share a similar structure, they do not fully overlap in student enrollment. Combining them would require restricting the sample to students enrolled in both courses or implementing nontrivial imputation strategies for missing observations. Either approach would substantially reduce the effective sample size and complicate interpretation. For these reasons, the analysis focuses exclusively on the Portuguese dataset.

2.2 Outcome Variable

The primary outcome variable is G3, the final course grade, measured on a standardized 0–20 scale. This variable captures cumulative academic performance at the end of the academic year and serves as the target for all predictive models. The focus on G3 aligns with the research question of understanding how behavioral, social, and family factors relate to final academic outcomes rather than short-term performance.

2.3 Predictor Domains

Rather than emphasizing individual preprocessing steps, predictors are conceptually organized into thematic domains that reflect different dimensions of student life and background:

- **Alcohol Consumption:** Measures of student alcohol use, summarized through an overall indicator of drinking behavior.
- **Academic Behavior:** Indicators related to study habits and academic history, including study time, absences, and prior failures.
- **Demographics:** Basic student characteristics such as age, sex, school attended, and home location (urban vs. rural).
- **Family Background:** Socioeconomic and household context, including parental education, parental occupations, family structure, and educational support at home.
- **Social and Lifestyle Factors:** Variables capturing peer interaction, free time, extracurricular activities, health status, and romantic relationships.

This domain-based organization clarifies how different types of predictors contribute to explaining variation in academic performance and helps structure the interpretation of later modeling results.

2.4 Modeling Readiness

All variables were retained in formats appropriate for regression-based and machine-learning methods, with categorical predictors prepared for indicator-based encoding and ordinal variables preserved in their natural order for interpretability. The dataset is fully observed, internally consistent, and structured to support both explanatory modeling and out-of-sample prediction in later sections of the analysis.

3 Data Cleaning and Preparation

3.1 Data Loading and Outcome Definition

The analysis uses the `student-por.csv` dataset from the UCI Machine Learning Repository (649 observations, 33 variables), chosen over the mathematics dataset for its larger sample size. Initial inspection confirmed no missing values and all variables within documented ranges. The primary outcome is G3 (final Portuguese grade). First- and second-period grades (G1, G2) were removed from predictors due to high correlations with G3 ($r > 0.8$ and 0.9 , respectively), which would introduce target leakage rather than reveal behavioral or family predictors.

3.2 Variable Typing and Encoding Decisions

All string-based variables were explicitly converted to categorical data types to reflect their qualitative nature and to prepare them for later encoding. Ordinal variables—such as `studytime`, `traveltime`, `Dalc`, `Walc`, `famrel`, `freetime`, `goout`, `health`, `Medu`, and `Fedu`—were intentionally kept as numeric rather than one-hot encoded. These variables have meaningful ordering, and treating them as numeric preserves interpretability of regression coefficients (e.g., moving from “low” to “high” consumption or education level).

Binary indicators (e.g., `schoolsup`, `famsup`, `paid`, `internet`, `romantic`) were retained in their original yes/no structure and later handled through one-hot encoding at the modeling stage. At the cleaning stage, the emphasis was on preserving semantic meaning rather than prematurely expanding the feature space.

3.3 Feature Engineering: Total Alcohol Consumption

The dataset includes two alcohol-related variables: weekday alcohol consumption (`Dalc`) and weekend alcohol consumption (`Walc`). Exploratory correlation analysis showed that these two measures are strongly correlated ($r \approx 0.62$), reflecting consistent drinking patterns across weekdays and weekends. To address redundancy and improve interpretability, a new feature—`Talc` (total alcohol consumption)—was constructed as the sum of `Dalc` and `Walc`.

This engineered variable provides a single, interpretable measure of overall alcohol use while reducing reliance on two highly overlapping predictors. Importantly, `Dalc` and `Walc` were not immediately dropped, allowing later comparisons between separate and aggregated alcohol measures during exploratory analysis and model selection.

3.4 Distribution Checks and Outlier Assessment

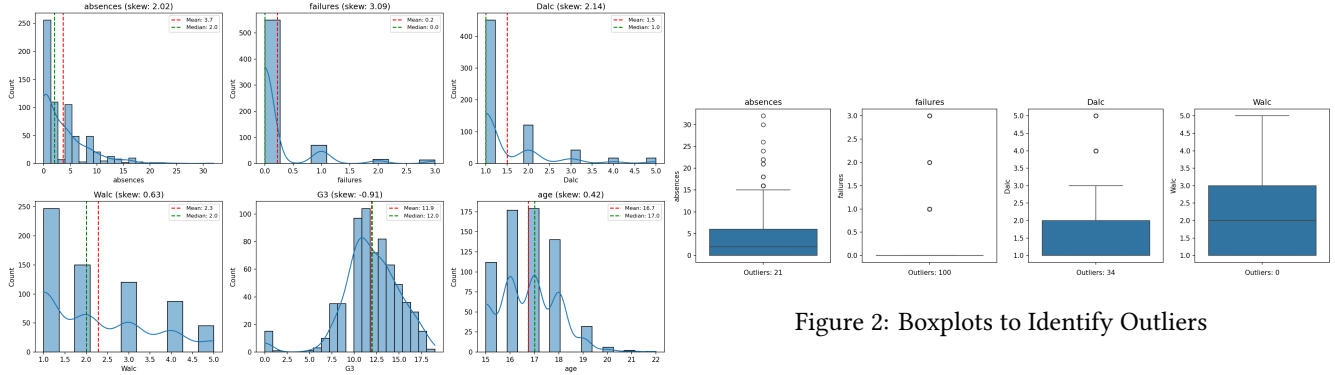


Figure 2: Boxplots to Identify Outliers

Figure 1: Distributions of Key Variables with Potential Outliers

Before modeling, distributions of all numeric predictors were examined using summary statistics, skewness measures, histograms, and boxplots. Several variables—such as failures, absences, and alcohol measures—exhibited strong right skew. These patterns reflect real behavioral phenomena (e.g., most students have zero failures or low alcohol consumption) rather than data errors.

No transformations were applied despite skewness. This decision was deliberate: the variables are bounded, interpretable on their original scales, and later modeling approaches (robust standard errors and penalized regression) are well-suited to handling non-normal predictors. Outliers identified via interquartile range methods were retained, as they represent genuine student experiences rather than noise. (Summary statistics for all numeric and categorical variables are provided in Appendix Tables A1 and A2.)

After removing leakage variables and adding the engineered Talc feature, the final cleaned dataset contains 649 observations and 32 variables. The cleaned dataset was saved for reproducibility and downstream modeling. At this stage, the data are internally consistent, free of leakage, and structured to support both interpretability-focused regression and more flexible predictive methods.

4 Exploratory Data Analysis (EDA)

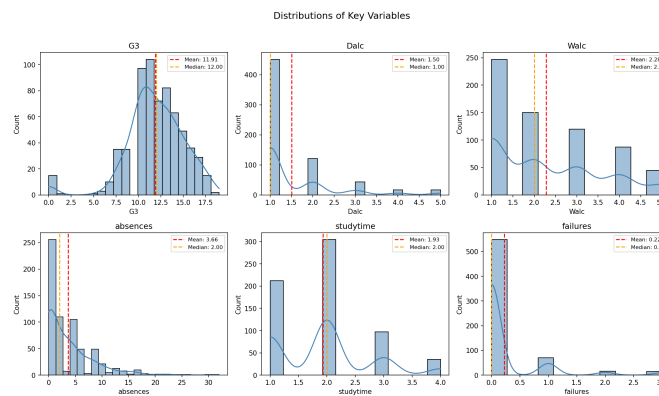


Figure 3: Distributions of Key Variables

4.1 Outcome Distribution

The final grade (G3) has a mean of approximately 11.9 (out of 20) and exhibits mild left skew, with most students scoring between 10 and 14. This distribution suggests meaningful variation in academic performance while avoiding extreme ceiling or floor effects, making it suitable for regression-based analysis.

4.2 Academic and Family Factors

Exploratory correlations highlight that academic history variables are much more strongly associated with final grades than alcohol consumption. The number of past failures is the single strongest correlate of G3, showing a substantial negative association. Study time is positively related to grades, while absences show a weaker negative relationship.

Family background variables, particularly parental education (Medu and Fedu), display moderate positive correlations with academic performance. These findings are consistent with socioeconomic explanations of educational outcomes and motivate their inclusion as controls in all subsequent models.

4.3 Social and Lifestyle Variables

Social variables such as goout and freetime exhibit weak negative correlations with grades, suggesting that increased social activity may be associated with slightly lower academic performance. However, these relationships are small in magnitude and likely intertwined with alcohol use and other behaviors, reinforcing the need for multivariate modeling.

4.4 Alcohol Consumption Patterns

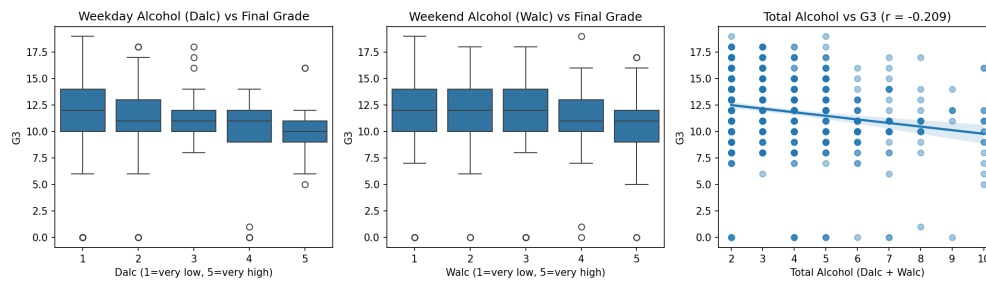


Figure 4: Alcohol Consumption vs Final Grade

Alcohol consumption is generally low in the sample. Weekday drinking (Dalc) is highly right-skewed, with most students reporting the minimum level, while weekend drinking (Walc) shows greater dispersion but remains concentrated at lower values. The strong correlation between Dalc and Walc confirms that these variables capture a shared underlying behavior rather than independent dimensions of alcohol use.

When examining alcohol variables against G3, all three measures (Dalc, Walc, and Talc) show weak but consistently negative correlations with final grades. Visualizations suggest a slight downward trend in grades as alcohol consumption increases, but the effect size is modest, indicating that alcohol use alone is unlikely to be a dominant predictor of academic performance.

4.5 Multicollinearity Assessment

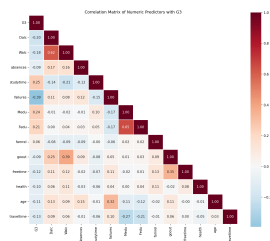


Figure 5: Correlation Matrix of Numeric Predictors with G3

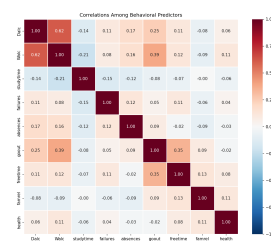


Figure 6: Correlations Among Behavioral Predictors

Correlation matrices among numeric predictors revealed notable clustering among behavioral variables, especially between Dalc and Walc and between parental education measures. Variance Inflation Factor (VIF) diagnostics showed that all VIF values were below conventional thresholds for severe multicollinearity, though alcohol variables had the highest values among predictors.

These results indicate moderate redundancy but no immediate need to remove variables purely on multicollinearity grounds. Instead, they motivate the use of regularization techniques and, as an exploratory exercise, Principal Components Analysis to understand latent structure in the data.

4.6 EDA Takeaways

Overall, EDA suggests that alcohol consumption is negatively related to academic performance but with small effect sizes, while academic history and educational aspirations dominate as predictors of final grades. Family socioeconomic indicators play a meaningful but secondary role, and behavioral and social variables are interrelated, creating moderate redundancy that must be handled carefully in modeling.

These insights directly inform the modeling strategy that follows, including the choice to compare baseline linear models, stepwise selection, and penalized regression approaches while carefully controlling for confounding factors.

5 Methodology

5.1 Principal Components Analysis (PCA)

To assess latent structure and redundancy among numeric predictors, Principal Components Analysis (PCA) was conducted as an exploratory dimensionality-reduction technique prior to predictive modeling. The goal of PCA in this analysis is not to replace original predictors in the main regression models, but to (i) diagnose multicollinearity, (ii) identify underlying behavioral and socioeconomic constructs, and (iii) inform later modeling choices.

5.1.1 PCA Inputs and Preprocessing

PCA was applied to 13 numeric predictors capturing demographics, academic behavior, social behavior, alcohol consumption, health, and attendance: age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences

The outcome variable (G3) was excluded, as were derived variables such as Talc, to avoid redundancy. All numeric predictors were standardized to mean 0 and standard deviation 1 using a StandardScaler, since PCA is sensitive to differences in scale.

5.1.2 PCA Estimation

PCA was fit using the full set of standardized numeric predictors. All components were initially retained to examine the full variance decomposition. Two diagnostic plots were generated:

- A scree plot showing variance explained by each component
- A cumulative variance plot showing the proportion of total variance explained as components are added

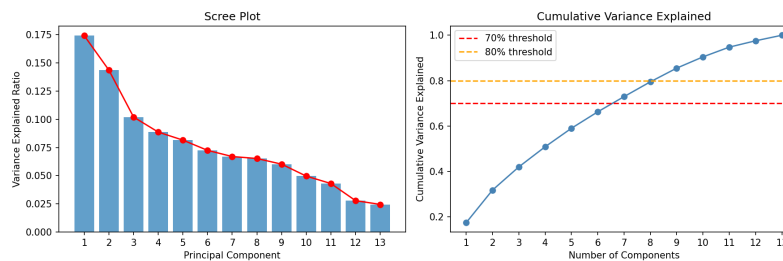


Figure 7: Scree Plot and Cumulative Variance Explained by PCA

To aid interpretation, component loadings for the first five principal components were examined and visualized using a heatmap. Loadings reflect the contribution of each original variable to a given component.

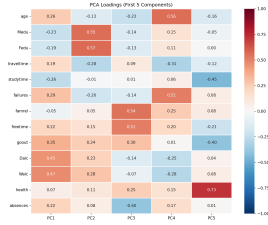


Figure 8: PCA Loadings for First Five Components

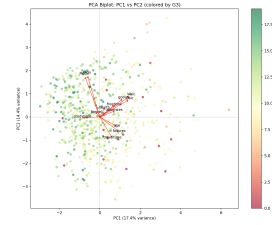


Figure 9: PCA Biplot: PC1 vs PC2 Colored by Final Grade

Finally, a biplot of the first two principal components was constructed, with observations projected into PC space and colored by final grade (G3). Variable loadings were overlaid as vectors to visualize how predictors align with the dominant axes of variation.

5.2 Linear, Stepwise, and Penalized Regression

5.2.1 Baseline Linear Regression

As a benchmark for more complex models, a baseline linear regression was estimated to quantify the relationship between alcohol consumption, academic behavior, family background, and final academic performance. This model provides a transparent reference point for assessing both explanatory power and predictive performance.

Train-Test Split: To evaluate out-of-sample performance and avoid overfitting, the cleaned dataset was split into training and test sets using an 80/20 split. The training set contains 519 students (80%), and the test set contains 130 students (20%). The split was performed using a fixed random seed to ensure reproducibility. Importantly, the mean and standard deviation of the outcome variable (G3) are similar across the training and test sets, indicating that the split preserves the overall grade distribution and does not introduce systematic imbalance. All preprocessing steps that could introduce leakage—most notably feature scaling—were fit exclusively on the training data and then applied to the test data.

Feature Encoding and Scaling The predictor matrix includes all variables except the outcome (G3) variable. Categorical predictors were converted to dummy variables using one-hot encoding with the first category dropped to avoid perfect collinearity. This resulted in a total of 40 predictors after encoding.

Numeric predictors were standardized to have mean 0 and standard deviation 1. Standardization facilitates comparison of coefficient magnitudes and improves numerical stability, especially given the heterogeneous scales of variables such as absences, age, and alcohol consumption. The scaler was fit on the training data only and then applied to both training and test sets, ensuring no information from the test set leaked into model training.

Model Specification An ordinary least squares (OLS) regression was fit on the training data using all available predictors. Given evidence of heteroskedasticity and non-normal residuals in exploratory diagnostics, heteroskedasticity-robust standard errors (HC3) were used for inference. This choice ensures more reliable hypothesis tests and confidence intervals without altering point estimates.

The baseline model includes alcohol consumption measures (Dalc, Walc, and the constructed Talc), academic behavior variables (e.g., failures, studytime, absences), family background indicators, demographic characteristics, and social/lifestyle factors. This “kitchen-sink” specification is intentionally broad, serving as a comprehensive reference against which more parsimonious or regularized models can be compared.

5.2.2 Stepwise Regression (AIC and BIC)

To balance model interpretability and predictive performance, stepwise linear regression was applied using information-criterion-based feature selection. Starting from an intercept-only model, predictors were added sequentially using forward selection. At each step, the candidate variable that produced the greatest improvement in model fit was selected, as measured by either the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

Both procedures were implemented on the training set only to avoid information leakage. All predictors were standardized prior to selection, and categorical variables were included through one-hot encoding. Two stepwise models were considered: An AIC-based model, which prioritizes predictive accuracy and allows a larger number of predictors. A BIC-based model, which applies a stronger penalty for model complexity and yields a more parsimonious specification.

The stepwise approach provides a contrast to the full baseline regression by explicitly removing redundant or weak predictors, allowing clearer assessment of whether alcohol-related variables remain important once irrelevant covariates are excluded.

5.2.3 Penalized Regression (Ridge, LASSO, Elastic Net)

Penalized regression methods were used to address multicollinearity, high dimensionality after one-hot encoding, and overfitting observed in the baseline linear model. These methods shrink coefficient magnitudes through regularization, improving stability and enabling variable selection.

The same training–test split (80/20) and preprocessing pipeline described earlier were applied. Predictors were standardized, categorical variables were one-hot encoded, and all preprocessing steps were fit on the training data only. The outcome variable is the final grade (G3), with intermediate grades excluded to prevent leakage.

Three models were estimated: First, a Ridge regression (L2 penalty), which shrinks correlated coefficients toward each other. Second, LASSO regression (L1 penalty), which performs shrinkage and variable selection by setting some coefficients to zero. Lastly the Elastic Net, which combines L1 and L2 penalties to balance sparsity and stability. The regularization parameter for each model was selected via cross-validation on the training set by minimizing mean squared error. Final models were refit using the optimal penalty and evaluated on the held-out test set.

5.3 Random Forest

After using linear regression to predict student grade as a continuous variable, we used random forest to classify their letter grade (A, B, C, etc.), and performance (high-performing and low-performing). In Portugal, universities use a 0-20 scale grading system, which can be translated to the letter grade system used in the US as follows.

Portugal Grade	US grade
18 - 20	A+
16 - 17.9	A
14 - 15.9	A-
12 - 13.9	B
10 - 11.9	C
0 - 9.9	F

Then, we will classify students whose grade is A+, A, and A- as high-performing students while those whose grade is B, C, and F as low-performing students.

To give our readers a brief overview of the random forest algorithm, random forest is an extension of the decision tree algorithm. A decision tree splits the dataset based on a binary condition of the data points' features recursively. For each split, among all possible conditions, the condition that decreases the variance of the dataset after split the most is selected. After a certain number of split or when the variance is under a desired threshold, the decision tree stops splitting. For any new data point, the prediction will be the mean of the datapoints that satisfy the same binary conditions of the decision tree.

A random forest algorithm creates multiple decision trees and aggregate the result of the trees to obtain the final result. It randomly selects a certain number of observations and a certain number of features or variables from the original dataset and then train a decision tree based on each selection. The aggregation of multiple decision trees based on randomly sampled data makes the model less sensitive to subtle changes in the training data.

After we trained the model, we will look at the top 20 important features the model used to classify student letter grade and performance. Also, we will present the confusion matrix and ROC curve to show where is the model failing at. We trained this model in python using the sklearn package.

5.4 XGBoost

Random forest is a very basic tree-based machine learning algorithm and we do not expect it to perform very well. Therefore, we also used XGBoost, a more sophisticated tree-based algorithm, to classify the letter grade and performance of students. We presented a detailed explanation of this algorithm in appendix 8.2 for readers who are interested to explore.

After we trained the model, we will look at the top 20 important features the model used to classify student letter grade and performance. Also, we will present the confusion matrix and ROC curve to show where is the model failing at.

We trained this model in python using the xgboost package.

5.5 Neural Networks & Text Mining

To explore non-linear predictive capabilities and extract qualitative context from categorical data, we employed two complementary approaches: neural network classification and text mining.

Text Mining: We aggregated multiple categorical variables—including parental jobs, school choice reasons, and support indicators—into a single tokenized text corpus. We then generated word clouds stratified by performance level to identify differences in the relative prominence of background factors between high-performing and low-performing students.

Neural Networks: We formulated a binary classification task to predict high performance (defined as $G3 \geq 14$, corresponding to A grades). We trained a Multi-Layer Perceptron (MLP) classifier with two hidden layers (64 and 32 units, ReLU activation) and compared it against a baseline decision tree (max depth 5). To monitor stability and overfitting, we tracked the training and validation log loss over 60 epochs.

6 Results

6.1 Principal Components Analysis

The PCA results indicate that variance is distributed across multiple dimensions rather than dominated by a single factor. The first principal component (PC1) explains approximately 17.4% of the total variance, followed by PC2 (14.4%) and PC3 (10.2%). The first five components together explain roughly 59% of total variance, while approximately 8–9 components are required to exceed 80% cumulative variance (Figure Figure 7).

This pattern suggests moderate redundancy among predictors but no extreme collinearity that would collapse the data into one or two dominant dimensions.

Inspection of loadings reveals interpretable latent constructs: The principal components reveal interpretable latent constructs. **PC1 (17.4% variance)** reflects a social–alcohol behavior dimension, driven by high loadings on Walc, Dalc, and goout. **PC2 (14.4% variance)** captures parental socioeconomic background through strong contributions from Medu and Fedu. **PC3 (10.2% variance)** represents a family and attendance dynamic, contrasting positive loadings on famrel and freetime with a negative loading on absences. **PC4 (8.9% variance)** reflects an academic risk profile associated with older age and prior failures, while **PC5 (8.2% variance)** distinguishes health status from academic effort, with high positive loading on health and negative loading on studytime and goout.

These patterns are visualized in the loading heatmap (Figure Figure 8) and biplot (Figure Figure 9).

Coloring the PCA biplot by G3 reveals no sharp separation of high- and low-performing students along PC1 or PC2. This suggests that while social/alcohol behavior and parental SES are meaningful dimensions of variation, they do not independently determine final grades. Instead, academic performance appears to vary along multiple axes simultaneously, reinforcing the need for multivariate regression rather than reliance on a small number of composite scores.

6.1.1 Implications for Modeling

The PCA results support three key decisions: 1. Behavioral and alcohol variables cluster together, confirming redundancy between Dalc, Walc, and related social measures. 2. Parental education forms a distinct latent construct, validating its treatment as a core control variable. 3. Despite moderate correlations, no severe multicollinearity is present, consistent with earlier VIF diagnostics.

Given these findings, PCA components are not used as primary predictors in the main models. Retaining original variables preserves interpretability for the research question, while later penalized regression methods handle redundancy more directly. PCA is therefore used as a diagnostic and interpretive tool rather than a replacement for the original feature set.

6.2 Linear, Stepwise, and Penalized Regression

Baseline Linear Regression. The baseline OLS model achieves an R^2 of 0.387 and adjusted R^2 of 0.338 with 40 predictors (F-test $p < 0.001$). None of the alcohol consumption measures are statistically significant once controls are included: Dalc ($\beta = -0.208$, $p = 0.302$), Walc ($\beta = +0.040$, $p = 0.808$), and Talc ($\beta = -0.073$, $p = 0.272$). This lack of significance reflects multicollinearity among alcohol variables and suggests alcohol consumption does not have a strong independent association with final grades. Academic history and aspirations dominate the results—past failures are the strongest negative predictor, while study effort and plans for higher education are positively associated with achievement (see Appendix Table A4 for significant predictors). These findings motivate stepwise selection and penalized regression to address multicollinearity.

Stepwise Regression. The AIC-based model retained 12 predictors, while the BIC-based model retained 7 (see Appendix Table A5). Both models retained failures, school_MS, higher_yes, Talc, schoolsup_yes, studytime, and health. Notably, total alcohol consumption (Talc) was retained in both models, whereas Dalc and Walc were excluded individually. Model

fit comparison shows the AIC model (Adjusted $R^2 \approx 0.345$) slightly outperformed the full baseline (Adjusted $R^2 \approx 0.338$) despite using far fewer predictors, while the BIC model (Adjusted $R^2 \approx 0.326$) achieved substantial parsimony with modest reduction in fit. In the AIC model, Talc shows a statistically significant negative coefficient—contrasting with the full baseline where alcohol variables were not significant due to multicollinearity. This supports the conclusion that alcohol consumption contributes modestly but non-negligibly to predicting academic performance when the model is restricted to informative predictors.

Penalized Regression. Figure Figure 10 shows the LASSO coefficient paths as a function of the regularization parameter α (log scale). As α increases, coefficients shrink toward zero, with the dashed red line indicating the cross-validated optimal value ($\alpha \approx 0.098$).

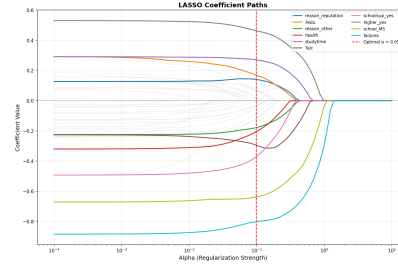


Figure 10: LASSO Coefficient Paths

The coefficient paths highlight a clear hierarchy of predictor importance. Academic difficulty (failures) remains large and negative across a wide range of α values and is among the last coefficients to be shrunk to zero, indicating strong and stable predictive power. School context and educational aspirations (school_MS, higher_yes) also persist under substantial regularization, reinforcing their central role in explaining final grades. Study behavior (studytime) shows a smaller but consistently positive contribution. Total alcohol consumption (Talc) is retained at the optimal penalty with a negative coefficient. Although its magnitude is modest relative to academic predictors, its persistence suggests that overall alcohol use contributes weakly but consistently to predicting lower grades once irrelevant variables are penalized away. In contrast, weekday and weekend alcohol measures (Dalc, Walc) shrink to zero early and are not retained at the optimal α , confirming redundancy among alcohol variables and supporting the use of a combined measure (see Appendix Table A3 for predictors retained across penalized models). Ridge regression retains all predictors but heavily shrinks correlated coefficients, while LASSO and Elastic Net exclude Dalc and Walc while retaining Talc, yielding a more parsimonious representation of alcohol use.

Figure ?@fig-penalized-performance compares test-set performance across penalized models.

All penalized models achieve comparable or slightly improved test RMSE relative to the unpenalized baseline, with reduced variance and greater stability. Differences across Ridge, LASSO, and Elastic Net are small, indicating that regularization primarily improves robustness rather than substantially increasing predictive accuracy. Overall, penalized regression confirms that academic history, study behavior, and educational aspirations dominate prediction of final grades. Alcohol consumption enters only through total consumption and with a comparatively small effect, indicating a secondary role once multicollinearity and overfitting are addressed.

6.2.1 Model Evaluation and Comparison

All models are evaluated on the same held-out test set using **test RMSE**, **test R^2** , and **test MAE**, with lower RMSE/MAE and higher R^2 indicating better generalization performance. Model complexity is measured by the number of retained features. The comparison focuses on how different modeling choices affect predictive performance and the accuracy–interpretability trade-off.

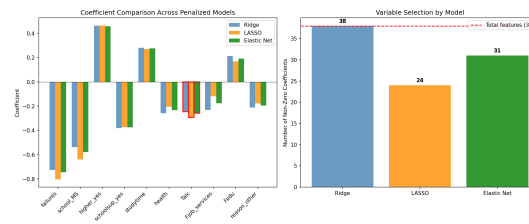


Figure 11: Model performance and complexity comparison

Table 8.1 reports the full set of performance metrics for each regression model.

Model	Type	Features	Train R^2	Test R^2	Test RMSE	Test MAE
OLS (Full)	Baseline	38	0.3863	0.1648	2.8538	2.1545
Stepwise (AIC)	Selection	12	0.3599	0.1512	2.8771	2.1570
Stepwise (BIC)	Selection	7	0.3350	0.1724	2.8409	2.1378
Ridge	Penalized	38	0.3780	0.2021	2.7895	2.0758
LASSO	Penalized	24	0.3616	0.1970	2.7983	2.0725
Elastic Net	Penalized	31	0.3715	0.2055	2.7836	2.0626

Across all models, test RMSE lies in a relatively narrow range (approximately 2.78–2.88), indicating that no single regression approach dramatically outperforms the others in predictive accuracy. Penalized models consistently achieve lower test RMSE and higher test R^2 than both the unpenalized OLS and stepwise models, suggesting that regularization improves generalization in the presence of correlated predictors. **Elastic Net achieves the lowest test RMSE (2.7836) and the highest test R^2 (0.2055), making it the best-performing model on the test set.**

To explicitly quantify the accuracy–interpretability trade-off, Table 8.2 reports test RMSE alongside feature counts, an interpretability score (higher values correspond to fewer features), and a combined trade-off score that balances predictive accuracy and interpretability.

Model	Test RMSE	Features	Interpretability	Trade-off
OLS (Full)	2.8538	38	0.26	0.17
Stepwise (AIC)	2.8771	12	7.11	3.55
Stepwise (BIC)	2.8409	7	8.42	4.27
Ridge	2.7895	38	0.26	0.28
LASSO	2.7983	24	3.95	2.11
Elastic Net	2.7836	31	2.11	1.22

The stepwise BIC model offers the highest interpretability by retaining only 7 predictors, but this comes with a modest reduction in predictive accuracy relative to penalized models. Ridge and Elastic Net prioritize predictive performance while retaining many predictors, making them less interpretable but more stable. LASSO occupies a middle ground, reducing dimensionality substantially while maintaining near-optimal test RMSE.

Overall, these results highlight a clear trade-off: **simpler models improve interpretability but do not yield the best predictive performance**, whereas **penalized models—especially Elastic Net—offer the strongest generalization at the cost of increased complexity**. This comparison provides the basis for interpreting the role of individual predictors, including alcohol consumption, in the context of model choice.

6.2.2 Implications for Alcohol Consumption (Talc)

To connect model comparison back to the research question, Talc’s role was examined across all regression models. Talc is retained in every model, and its standardized coefficient is consistently negative.

Model	Talc Coefficient
OLS (Full)	-0.2264
Stepwise (AIC)	-0.3717
Stepwise (BIC)	-0.4458
Ridge	-0.2455
LASSO	-0.2945
Elastic Net	-0.2585

The direction of the Talc coefficient is stable across modeling strategies, indicating a robust negative association between total alcohol consumption and final grades after controlling for academic and contextual factors. However, the magnitude of Talc’s effect is small relative to dominant predictors such as prior failures, study time, and educational aspirations.

Importantly, although Talc is consistently selected, **including alcohol consumption does not materially change overall predictive performance**. Differences in RMSE across models are driven primarily by regularization and feature

selection rather than by whether alcohol is included. This suggests that alcohol consumption contributes a modest, secondary signal rather than serving as a primary driver of academic performance.

Overall, the model evaluation supports a nuanced conclusion: alcohol consumption is not irrelevant—its negative association with grades is stable across models—but its role is limited compared to core academic factors, and it does not substantially improve prediction on its own.

6.3 Random Forest

6.3.1 Predicting Student Letter Grade

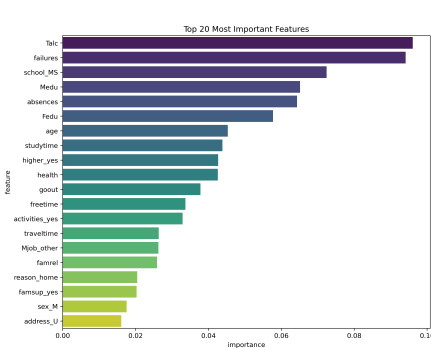


Figure 12: Top 20 Important Features

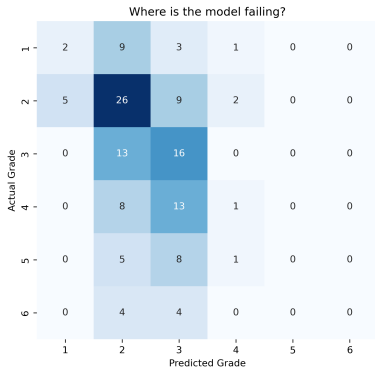


Figure 13: Confusion Matrix

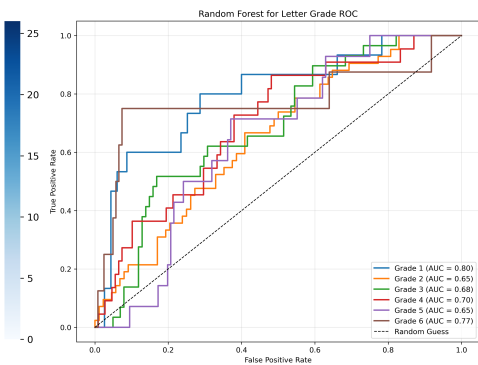


Figure 14: ROC Curve

As we can see in Figure 12, total alcohol consumption is the most important feature in predicting a student’s letter grade. It is slightly more important than the number of classes failed in the past and significantly more important than any other features. This result strongly confirms our hypothesis that total alcohol consumption contributes well to the prediction of students’ grade.

However, the test error rate of random forest is 0.6538, meaning the model predicted 65.38% of the data wrong in the test set. But note that we have 6 levels here (A+, A, A-, B, C, F), the error rate of random guessing is 0.8333. Therefore, although the test error rate is not very low, it is still much lower than random guessing, making our model meaningful in predicting student letter grade.

Now, we present the confusion matrix to see where is the model failing at. Note that 1 - 6 corresponds to F - A+ respectively.

As we can see in Figure 13, the model made no prediction of A or A+ and only 5 predictions of A-. This is because the number of students scoring higher or equal to A- takes up only 33% of the test set. Thus, the random forest model learns that it can perform better by just making less or no prediction of A-, A, and A+. After realizing this, we modified the code to force the model to make predictions of A- to A+. But the resulting test error rate is even higher. Therefore, such bias might just be an inherent flaw in the algorithm of random forest.

Finally, we present the ROC curve for this model.

As we can see in Figure 14, the model performs the best when predicting F and A+ as their curves have the highest AUC score.

6.3.2 Predicting Student Performance

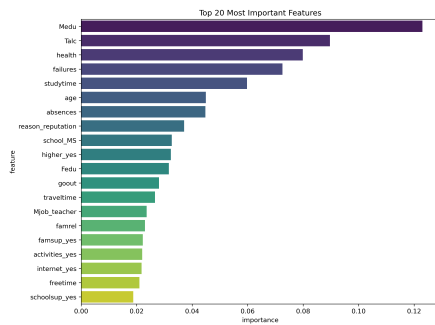


Figure 15: Top 20 Important Features

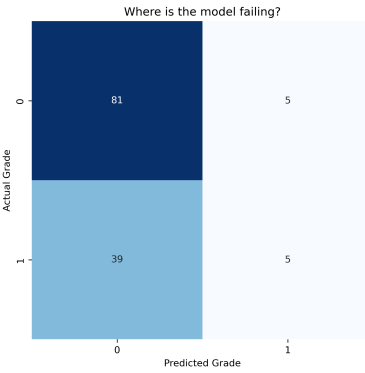


Figure 16: Confusion Matrix

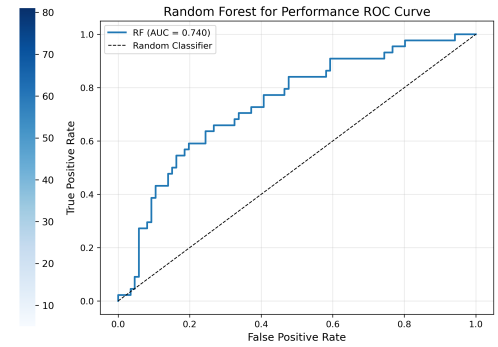


Figure 17: ROC Curve

As we can see in Figure 15, when predicting student performance (high or low), the most important feature is the education of the mother, which significantly outperformed all other predictors, followed by total alcohol consumption. The number of classes failed, which is the second important feature in predicting student letter grade, now ranks the fourth. Therefore, we can still conclude that total alcohol consumption is an important predictor of student performance.

The test error rate for this model is 0.3385, meaning that it predicted 33.85% of the data wrong in test set. Note that we cannot compare this error rate with the previous one as the outcome variable here, the performance, is binary, making random guessing having a error rate of 50%.

The test error rate is still pretty high. And we present the confusion matrix to see where is the model failing at. Note that 0 is for low performance and 1 is for high performance.

As we can see in Figure 16, we encountered the same problem as in 4.2.1. As there are fewer students with high performance, the model learns that it can perform better by making less prediction of high performance.

Finally, we present the ROC curve for this model.

From Figure 17, we can see that the ROC curve is moderately higher than the diagonal, meaning that the model is moderately better than random guessing, but still not very good at predicting performance.

6.4 XGBoost

6.4.1 Predicting Student Letter Grade

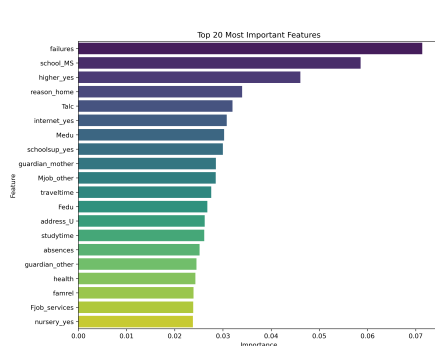


Figure 18: Top 20 Important Features

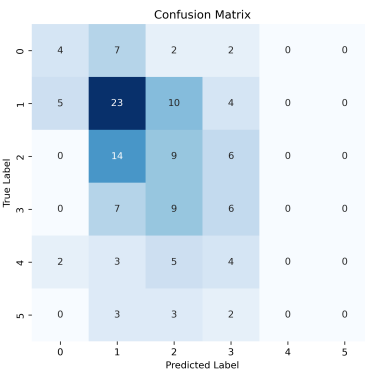


Figure 19: Confusion Matrix

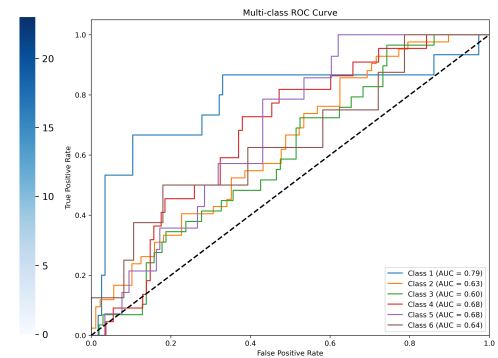


Figure 20: ROC curve

As shown in Figure 18, among the top 20 most important features to predict student letter grade by XGBoost, total alcohol consumption ranks the fifth. The most important feature is the number of classes failed, which is also considered to

be very important by random forest. We can conclude that according to XGBoost, total alcohol consumption is an important feature to predict student letter grade.

The test error rate for this model is 0.6769, meaning that the model predicted 67.69% of data wrong in the test set. Comparing to the test error rate of 0.6538 by random forest on the same problem, XGBoost does not exhibit much improvement in performance.

Now, we present the confusion matrix to see where is the model failing at.

From Figure 19, we can see the same problem in Figure 13: the model learns that it can perform better by just making less or no prediction of A-, A, and A+. Since this problem is not unique to random forest, it might be a common flaw in tree-based algorithms.

From Figure 20, we can see that the model is best at predicting grade F, which is consistent with the confusion matrix.

6.4.2 Predicting Student Performance

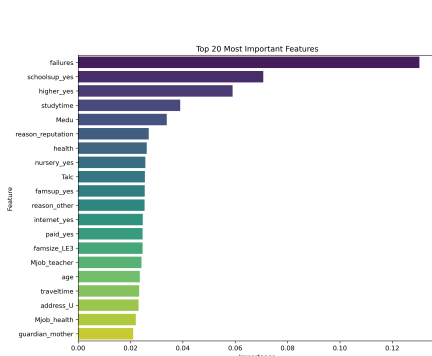


Figure 21: Top 20 Important Features

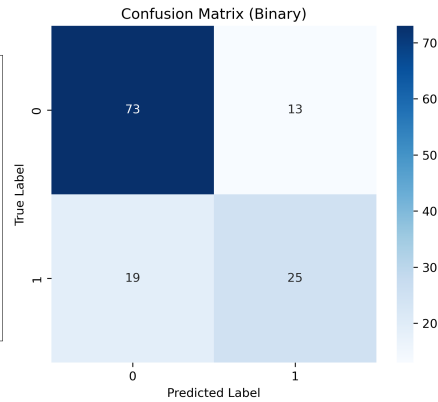


Figure 22: Confusion Matrix

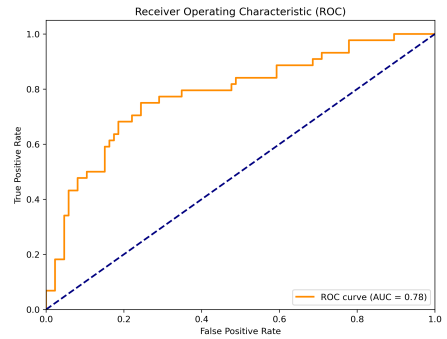


Figure 23: ROC curve

From Figure 21, we can see that the number of classes failed significantly outranks all other features to be the most important one while total alcohol consumption ranks only the ninth. Considering there are a total of 38 features, ranking the ninth still implies total alcohol consumption to be fairly important in predicting student letter grade.

The test error rate is 0.2462, meaning that it predicted 24.62% of the data wrong in the test set. Comparing to the test error rate of 0.3385 by random forest on the same problem, XGBoost exhibits great improvement in performance.

From the confusion matrix, we can see that the model is not just predicting less high performance (level 1) to improve performance. Although it is still biased towards low performance, it is a lot better than the random forest model as shown in Figure 16.

Compared to the AUC of 0.74 by the random forest model, the XGBoost does not show much improvement with an AUC of 0.78.

6.5 Neural Networks & Text Mining Results

6.5.1 Qualitative Insights from Text Mining

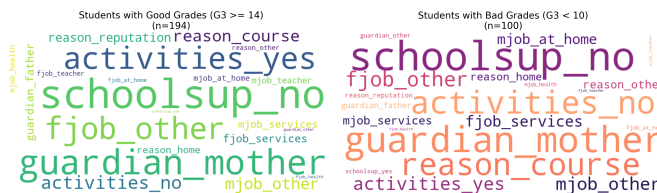


Figure 24: Word Clouds of Student Backgrounds by Performance Level

Because the corpus is built from `column_value` tokens (coded categorical levels such as `guardian_mother` and parental job labels), the word clouds should be interpreted as a frequency summary of background categories rather than natural-language

themes. To keep definitions consistent with our classification task, we stratified students by high performance ($G3 \geq 14$) versus not high performance ($G3 < 14$). The two clouds share many dominant background tokens, suggesting that broad family context is relatively similar across groups, while any differences are more subtle and engagement-related (for example, extracurricular participation indicators like `activities__yes` vs `activities__no`). Overall, the word clouds are descriptive context rather than strong evidence of causal drivers.

6.5.2 Classification Performance (Neural Network vs Tree)

We compared the Multi-Layer Perceptron (MLP) against a standard Decision Tree to test whether non-linear complexity improves prediction for high-performing students. On the held-out test set, both the Neural Network and the Decision Tree achieved an accuracy of **0.7000**, which is approximately the majority-class baseline in this split. This suggests that, under the $G3 \geq 14$ threshold and without prior grade history (G1, G2), the available demographic and background variables provide limited signal for distinguishing top-performing students. For this reason, accuracy alone is not sufficient, and confusion-matrix and ROC-style diagnostics are needed to evaluate minority-class performance.

6.5.3 Training Dynamics

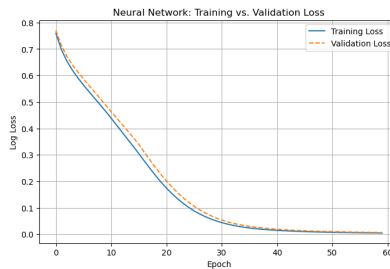


Figure 25: Neural Network Training vs Validation Loss

The loss curve shows validation loss flattening (and even slightly increasing) while training loss continues to decrease. This divergence indicates that the model began overfitting the training noise relatively quickly. The inability of the validation loss to decrease significantly below its starting point confirms that the demographic features lack strong signals for this specific high-performance threshold ($G3 \geq 14$).

6.5.4 Feature Importance (Bagged Tree)

To verify the role of alcohol, we analyzed the feature importance from the bagged tree model.

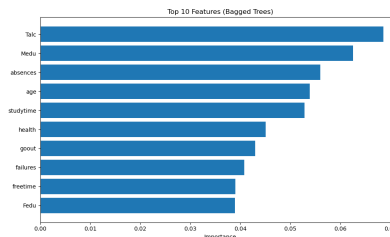


Figure 26: Top 10 Features (Bagged Tree)

As shown in Figure Figure 26, alcohol consumption (Talc) does not appear in the top 10 predictors, confirming it is a weak signal compared to failures, absences, and parental education.

7 Conclusion

7.1 Linear, Stepwise, and Penalized Regression

Across all regression approaches, three consistent findings emerge. First, academic history and behavior—particularly past failures, study time, and educational aspirations—are the dominant predictors of final grade. These variables are statistically

strong, stable across model specifications, and central to out-of-sample predictive accuracy. Second, model comparison reinforces that no single regression model drastically outperforms the rest in predictive performance; test RMSE values fall within a narrow range (~2.78–2.88), with Elastic Net performing best by a small margin. Penalized regression improves robustness and generalization relative to the full OLS baseline, while stepwise models improve interpretability by removing redundant predictors.

Finally, alcohol consumption shows a consistent but secondary signal. Total consumption (Talc) is retained across nearly all models, its coefficient remains negative, and its effect persists after accounting for academic, family, demographic, and behavioral controls. However, effect sizes are small relative to core academic predictors, and including alcohol variables does not meaningfully change overall predictive accuracy. Taken together, the regression results suggest that alcohol matters—but as a modest contributor alongside more powerful academic drivers rather than as a primary determinant of performance.

7.2 Tree Based Models

We used random forest and XGboost to predict the letter grade and the performance of students. When predicting letter grade, both random forest and XGboost consider total alcohol consumption to be one of the most important features in prediction. Their test error rate and AUC scores when predicting each grade are very similar to each other, indicating similar and moderately good model performance.

When predicting student performance, random forest consider total alcohol consumption to be the second most important feature while XGBoost only consider it to be the ninth important feature. XGBoost outperformed random forest significantly but their AUC scores are very similar. This indicates that the discriminative power of both models are similar, but differ in their calibration. Future research could try to finetune the random forest model to see if it can perform as good as XGBoost.

Overall, we can conclude that total alcohol consumption is one of the most important predictors of student letter grade and performance according to random forest and XGBoost.

For limitations, both models are biased towards the class with more observations (lower letter grade and low performance) and forcing them to balance the prediction would increase test error rate, which is probably the shared flaw of tree-based classification algorithms. Therefore, tree-based classification algorithms might not be the best technique when the dataset is unbalanced. In addition, we used random forest and XGBoost for classification, making it impossible to compare their performance with our linear regression model. Since random forest and XGBoost can also be applied in regression (although it is less common), future research can use the regression version of the two algorithms and compare their performance with that of linear regression.

7.3 Neural Networks & Text Mining

Overall, the neural network performed identically to the shallow decision tree in this split, with neither model managing to outperform the majority-class baseline. The loss dynamics highlight that while the network began to learn, it quickly plateaued, suggesting that demographic variables alone are insufficient to reliably identify “A-student” performance. Since the neural network did not demonstrate superior accuracy here, its “black box” nature makes it less attractive than transparent models for this specific task.

7.4 Limitations and Future Directions

Several limitations constrain interpretation. The observational, cross-sectional design precludes causal inference—alcohol’s association with lower grades may reflect unmeasured confounders such as stress or disengagement. Self-reported predictors (alcohol use, study time, health) introduce potential bias, and the single-subject outcome may be influenced by instructor-specific factors. Modest predictive accuracy (test $R^2 \approx 0.20$) suggests key determinants remain unmeasured.

Future work could apply this pipeline to the mathematics dataset to test whether alcohol’s association with performance generalizes across subjects. Extensions incorporating longitudinal designs, interaction effects, or richer behavioral measures may improve predictive power. Regarding alcohol specifically, Talc shows a small but consistent negative association with grades across models, though its effect is modest relative to academic predictors. The evidence does not support alcohol as a standalone screening indicator; rather, it should be treated as one component within a broader academic-risk profile.

8 Appendix

8.1 Summary Statistics Tables and Selected Predictors Tables

Table A1. Summary Statistics for Numeric Variables

Statistic	age	Medu	Fedu	travel-time	study-time	fail-ures	fam-rel	free-time	goout	Dalc	Walc	health	ab-sences	G3	Talc
count	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00	649.00
mean	16.74	2.51	2.31	1.57	1.93	0.22	3.93	3.18	3.18	1.50	2.28	3.54	3.66	11.91	3.78
std	1.22	1.13	1.10	0.75	0.83	0.59	0.96	1.05	1.18	0.92	1.28	1.45	4.64	3.23	1.99
min	15.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	2.00
25%	16.00	2.00	1.00	1.00	1.00	0.00	4.00	3.00	2.00	1.00	1.00	2.00	0.00	10.00	2.00
50%	17.00	2.00	2.00	1.00	2.00	0.00	4.00	3.00	3.00	1.00	2.00	4.00	2.00	12.00	3.00
75%	18.00	4.00	3.00	2.00	2.00	0.00	5.00	4.00	4.00	2.00	3.00	5.00	6.00	14.00	5.00
max	22.00	4.00	4.00	4.00	4.00	3.00	5.00	5.00	5.00	5.00	5.00	5.00	32.00	19.00	10.00

Table A2. Summary Statistics for Categorical Variables

Variable	Count	Unique	Most Common Category	Frequency
school	649	2	GP	423
sex	649	2	F	383
address	649	2	U	452
famsize	649	2	GT3	457
Pstatus	649	2	T	569
Mjob	649	5	other	258
Fjob	649	5	other	367
reason	649	4	course	285
guardian	649	3	mother	455
schoolsup	649	2	no	581
famsup	649	2	yes	398
paid	649	2	no	610
activities	649	2	no	334
nursery	649	2	yes	521
higher	649	2	yes	580
internet	649	2	yes	498
romantic	649	2	no	410

Table A3. Predictors Retained at Optimal Penalty Across Penalized Models

Predictor	Ridge	LASSO	Elastic Net
failures	☒	☒	☒
studytime	☒	☒	☒
higher_yes	☒	☒	☒
school_MS	☒	☒	☒
schoolsup_yes	☒	☒	☒
health	☒	☒	☒
Medu / Fedu	☒	☒	☒
Talc	☒	☒	☒
Dalc	☒	☒	☒
Walc	☒	☒	☒

Table A4. Significant Predictors of Final Grade (Baseline Regression, $p < 0.05$)

Variable	Coefficient	p-value	Interpretation
failures	-0.884	<0.001	Each SD increase in past failures is associated with a 0.88-point decrease in final grade
higher_yes	+1.687	<0.001	Aspiring to higher education is associated with a 1.7-point increase in final grade
schoolsup_yes	-1.603	<0.001	Receiving school support is associated with a 1.6-point decrease (likely reverse causation)
school_MS	-1.394	<0.001	MS school students score about 1.4 points lower than GP students
Fjob_services	-1.099	0.030	Father working in services is associated with a 1.1-point decrease relative to at_home
studytime	+0.300	0.011	Each SD increase in study time corresponds to a 0.3-point increase in final grade
health	-0.328	0.013	Better self-reported health is associated with lower grades (likely confounded)

Table A5. Selected Predictors Under AIC and BIC Stepwise Regression

Variable	AIC Model	BIC Model
failures	☒	☒
school_MS	☒	☒
higher_yes	☒	☒
Talc	☒	☒
schoolsup_yes	☒	☒
studytime	☒	☒
health	☒	☒
Fedu	☒	
reason_other	☒	
absences	☒	
Fjob_services	☒	
reason_reputation	☒	

8.2 Explanation of XGBoost Algorithm

To understand XGBoost, we must first understand gradient boosting. Gradient boosting is an extension of the decision tree algorithm. It creates multiple decision trees and aggregate the results to obtain the final result. But unlike random forest, every decision tree gradient boosting creates depends on the results of the previous tree. The learning process starts from simply taking the mean of the dependent variable in the training set. Then, the residual error for each observation is calculated and a new decision tree is built to split the data until a certain number of leaves, which is preset as a parameter, is created. As the number of leaves is restricted, if more than one value is on the same leave, the value on that leaf will be replaced with their mean.

The predictions of residuals obtained by a decision tree will be scaled by a factor between 0 to 1, named learning rate, to prevent the model from over-fitting to the training data. And the actual prediction of this new tree will be the prediction of the previous tree plus the product of learning rate and the predicted residuals from the new tree. Then, the new residual error is calculated and the same process is repeated until a certain number of decision trees, which is also a preset parameter, is created. The final result \hat{Y} of gradient boosting is as follows:

$$\hat{Y} = \bar{Y} + \gamma r_1 + \gamma r_2 + \cdots + \gamma r_n$$

Where \hat{Y} is the mean of training data, r_i is the residual error predicted by each decision tree, and γ is the learning rate.

Now, we move on to XGBoost. XGBoost is an extension of the Gradient Boosting algorithm. Similar to Gradient Boosting, it creates multiple trees to predict the residual error made by the past prediction and aggregate them to produce the final result. The differences between the two algorithms are the condition of splitting data and create branches in the decision tree, the inclusion of a regularization parameter, and the inclusion of a pruning process.

In Gradient Boosting, the condition to split data at each node in the decision tree is the one that decreases the variance of the data the most after split. While in XGBoost, the condition is the one that increases the similarity score of the data the most after split. The algorithm stops splitting the data after a certain number of times, which is a preset parameter. The similarity score of a leaf is calculated as:

$$\text{Similarity Score} = \frac{(\sum \text{Residual})^2}{\text{Number of Residuals} + \lambda}$$

Where γ is a preset regularization parameter, ranging from 0 to positive infinity, that prevents the model from overfitting to the training data. After the split, the condition that maximizes the increase in similarity score is chosen. This parameter is also included when calculating the prediction result of each leaf:

$$\text{predicted residual} = \frac{\sum \text{residual}}{\text{number of residual} + \lambda}$$

Before the algorithm move on to build the next tree, XGBoost has an additional step of pruning. Starting from the last leaf created in the decision tree, if the increase in similarity score is smaller than gamma, which is a preset parameter, that split will be pruned, or eliminated from the decision tree. The process continues to the second last leaf in the decision tree until the increase in similarity score of one leaf is larger than gamma.

Then, next tree is built based on the residual error of the prediction of this tree until a certain number of trees has been created. The final result is also calculated the same way as in Gradient Boosting: summing the base of prediction and all the products of predicted residuals and a learning rate.