Predicting Student Alcohol Consumption

Naina Sharma, Ava Adlao, Ray Min, Jinhyo Lee, Michael Gureghian

STATS 112

Dr. Mahtash Esfandiari

December 9, 2024

Table of Contents

1. Abstract

The objective of this study was to observe the influence of 28 demographic, academic performance, educational support, and social factors in predicting secondary school students' level of alcohol consumption given a 2008 survey of 1044 students.

Logistic regression using a reduced model with 10 predictors indicated the odds of moderate-to-high alcohol use is increased significantly in male students and for students who indicated "other" when asked why they chose to attend their school; it also has a significant positive correlation with number of absences and how often a student goes out with friends. This odds decreases significantly for students whose mothers work in health, whose primary guardian is their mother, and with increased study-time, higher grades, and better-quality familial relationships.

The training accuracy of this regression was evaluated at 0.7634 with a confusion matrix and 0.7686 with AUC.

2. Introduction

Adolescent alcohol consumption is a significant public health concern, with long-term implications for health, academic achievement, and social well-being. Understanding the predictors of alcohol use among secondary school students is crucial for developing targeted interventions. This study utilized demographic, academic, social, and familial data to analyze patterns in weekday alcohol consumption among the secondary school students. By applying the statistical methods such as logistic regression, we aim to identify key factors influencing alcohol use and provide actionable insights.

i. Data

The dataset used in our study was obtained from Kaggle and originates from a 2008 survey of Portuguese secondary school students. It includes data on 1,044 students enrolled in Math and Portuguese language courses across two schools. The dataset consists of 30 variables categorized into demographic, academic performance, social relationships, and educational support groups. The primary outcome variables are weekday alcohol consumption (Dalc) and weekend alcohol consumption (Walc), with a focus on Dalc for this analysis.

ii. Variables and Path diagram

The dataset comprises 30 variables, which can be categorized as follows:

Demographic Variables

- school: Indicates the student's school. Binary variable with two levels: GP (Gabriel Pereira) or MS (Mousinho da Silveira).
- sex (gender): Indicates the student's gender. Binary variable: F (Female) or M (Male).

- age (Age): Represents the student's age in years. Ordinal variable ranging from 15 to 22.
- address (Address type): Indicates the type of area where the student lives. Binary variable: urban or rural.
- family size (Family size): Represents the size of the student's family. Binary variable: 3 (small family) or >3 (large family).
- parents' cohabitation status: Indicates whether the student's parents live together. Binary variable: living together or apart.
- Medu (Mother's education level): Represents the education level of the student's mother. Ordinal variable ranging from 0 (none) to 4 (higher education).
- Fedu (Father's education level): Represents the education level of the student's father. Ordinal variable ranging from 0 (none) to 4 (higher education).
- Mjob (Mother's job): Describes the type of job the student's mother has. Categorical variable with levels: teacher, health, services, at_home, other.
- Fjob (Father's job): Describes the type of job the student's father has. Categorical variable with levels: teacher, health, services, at_home, other.
- guardian: Indicates the student's primary guardian. Categorical variable with levels: mother, father, other.

Academic Performance Variables

- grades: Represents the average of the student's grades over three periods. Numeric variable ranging from 0 to 20.
- health: Self-reported health status of the student. Ordinal variable ranging from 1 (very poor) to 5 (excellent).
- absences: The total number of school absences. Numeric variable.

Social Relationship Variables

- activities (Extracurricular activities): Indicates if the student participates in extracurricular activities. Binary variable: yes or no.
- nursery (Nursery attendance): Indicates if the student attended nursery school. Binary variable: yes or no.
- higher (Aspiration for higher education): Indicates if the student plans to pursue higher education. Binary variable: yes or no.
- internet(Internet access): Indicates if the student has internet access at home. Binary variable: yes or no.
- romantic: Indicates if the student is in a romantic relationship. Binary variable: yes or no.
- gout : Represents how often the student goes out with friends. Ordinal variable ranging from 1 (very low) to 5 (very high).
- famrel: Self-reported quality of relationships within the family. Ordinal variable ranging from 1 (very poor) to 5 (excellent).
- freetime: Represents how much free time the student has after school. Ordinal variable ranging from 1 (very low) to 5 (very high).

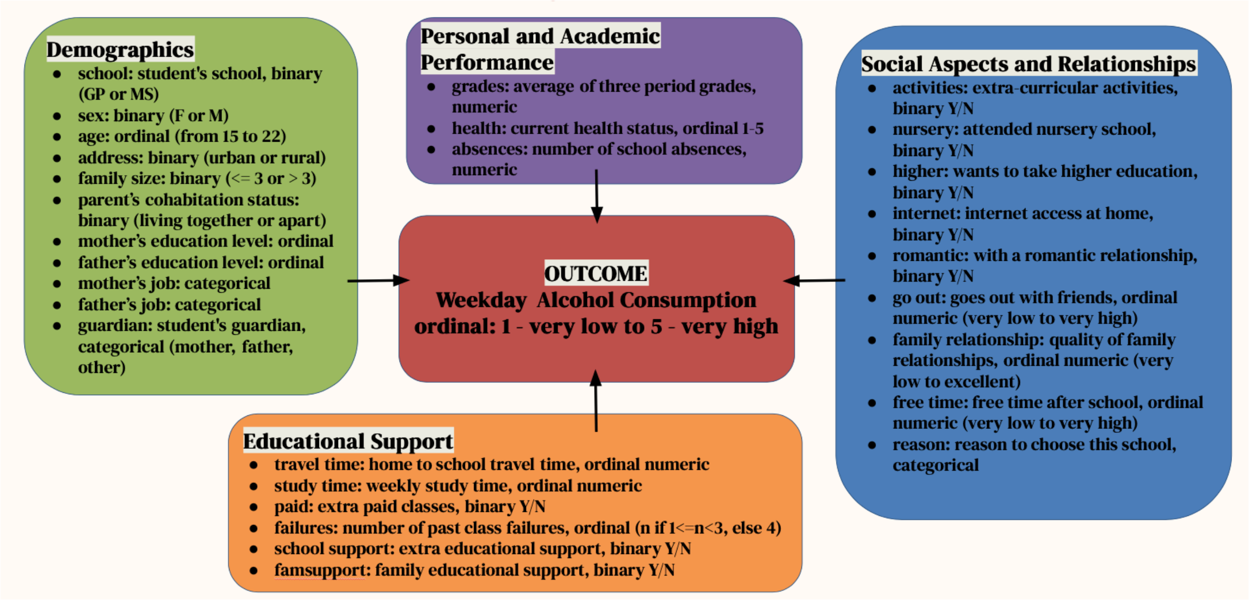
Educational Support Variables

- traveltime: Indicates the time it takes for the student to travel from home to school. Ordinal variable ranging from 1 (<15 minutes) to 4 (>1 hour).
- studytime: Indicates the weekly study time of the student. Ordinal variable ranging from 1 (<2 hours) to 4 (>10 hours).
- failures: Represents the number of past class failures. Ordinal variable: 0 (none), 1, 2, or 3 (failures).
- paid: Indicates if the student receives extra paid classes. Binary variable: yes or no.
- schoolsup: Indicates if the student receives extra educational support from the school. Binary variable: yes or no.
- famsup) Indicates if the student receives educational support from the family. Binary variable: yes or no.

Outcome Variable:

- Dalc(Weekday alcohol consumption): Ordinal variable (1 - very low to 5 - very high). It indicates students' alcohol consumption during weekdays. In addition, we change binary variable because Dalc distribution is imbalanced. Binary 0: Low alcohol consumption (1) or 1: Higher alcohol consumption (2-5).
- Walc(Weekend alcohol consumption): Ordinal variable (1 - very low to 5 - very high). It indicates students' alcohol consumption during weekdays. In addition, we change binary variable because Dalc distribution is imbalanced.

Path diagram:


Demographics
● school: student's school, binary (GP or MS)
● sex: binary (F or M)
● age: ordinal (from 15 to 22)
● address: binary (urban or rural)
● family size: binary (<= 3 or > 3)
● parent's cohabitation status: binary (living together or apart)
● mother's education level: ordinal
● father's education level: ordinal
● mother's job: categorical
● father's job: categorical
● guardian: student's guardian, categorical (mother, father, other)
Personal and Academic Performance
● grades: average of three period grades, numeric
● health: current health status, ordinal 1-5
● absences: number of school absences, numeric
Social Aspects and Relationships
● activities: extra-curricular activities, binary Y/N
● nursery: attended nursery school, binary Y/N
● higher: wants to take higher education, binary Y/N
● internet: internet access at home, binary Y/N
● romantic: with a romantic relationship, binary Y/N
● go out: goes out with friends, ordinal numeric (very low to very high)
● family relationship: quality of family relationships, ordinal numeric (very low to excellent)
● free time: free time after school, ordinal numeric (very low to very high)
● reason: reason to choose this school, categorical
OUTCOME
Weekday Alcohol Consumption
ordinal: 1 - very low to 5 - very high
Educational Support
● travel time: home to school travel time, ordinal numeric
● study time: weekly study time, ordinal numeric
● paid: extra paid classes, binary Y/N
● failures: number of past class failures, ordinal (n if 1<=n<3, else 4)
● school support: extra educational support, binary Y/N
● famsupport: family educational support, binary Y/N

iii. Research Question

Does demographics, personal and academic performance, social aspects and relationships, and educational support predict the weekday alcohol consumption of a student in secondary school?

3. Literature Review

Alcohol is a leading cause of death for people between the ages of 15 and 29 (World Health Organization, 2018). Past research has used survey data with categorical, ordinal, and numeric predictors to study student alcohol consumption represented with binary and ordinal response variables.

Posavec (2022) used a logistic regression to study and assess the responses to a questionnaire he distributed to 1352 Varaždin third-year high school students aged 15-20. This survey investigated primarily psychosocial factors and students' use of electronic devices in relation to their alcohol use. Posavec used chi-square tests and t-tests to assess feature significance; he concluded that parenting style and maternal authority were the most important factors towards decreasing likelihood of alcohol consumption, and that a student's use of electronic media had no significant correlation with alcohol consumption (2022). He also found that whether or not a student's father drinks has an extremely large impact on the student's alcohol consumption.

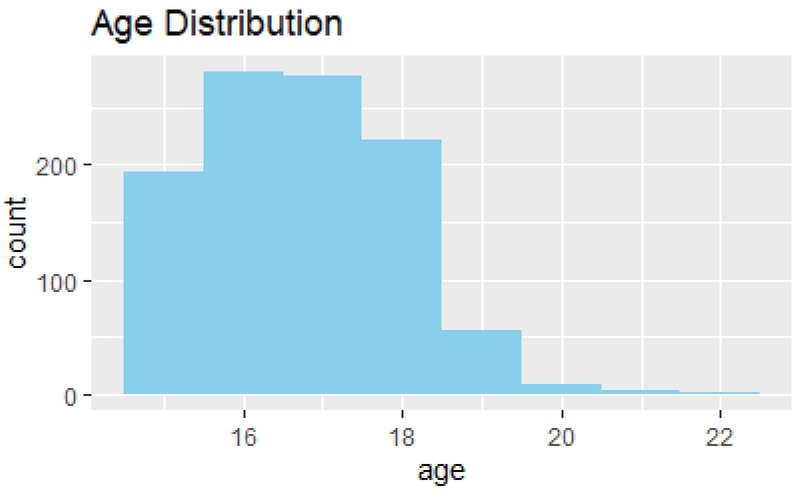
Esser et al. (2017) studied rates of binge and active drinking in American high school students from 1991-2015 using sociodemographic predictors obtained from a yearly survey of high schools across the U.S. Using time as a continuous variable, they applied a logistic regression and found that overall drinking had declined significantly from 1991 to 2015. They also found that although female students' alcohol consumption was lower than male students' in 1991, the rate at which their alcohol consumption decreased by 2015 was lower than that of male students, so the rates of male and female students' alcohol consumption converged (2017).

Balsa et al. (2012) used survey data with a very large sample size of American high school students in the 1994/1995 school year to investigate the impact of alcohol consumption on GPA and academic performance. Using linear regressions separated by gender, they came to the conclusion that alcohol consumption has a significant yet small impact on GPA for male students and an insignificant effect for female students, and that for female students, higher rates of alcohol consumption correlate with a higher feeling of academic difficulty (2012).

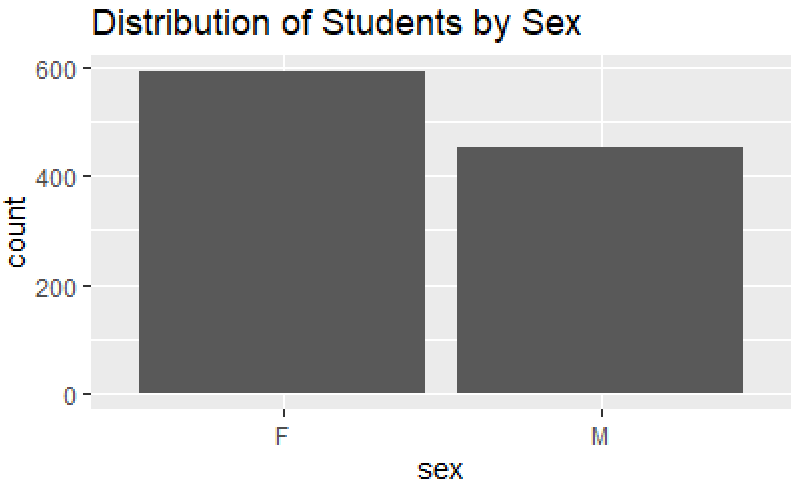
4. Exploratory Data Analysis

This report explores the factors influencing alcohol consumption among secondary school students based on survey data from Portuguese students enrolled in Math and Portuguese courses. The dataset, sourced from Kaggle, includes demographic, academic, social, and support-related variables. The primary response variable is Dalc, which measures weekday alcohol consumption on a five-point ordinal scale. The study's objective is to identify significant predictors of alcohol consumption and assess how different factors influence drinking behavior among adolescents. Before the models were created, initial analysis and graphics of the variables were created.

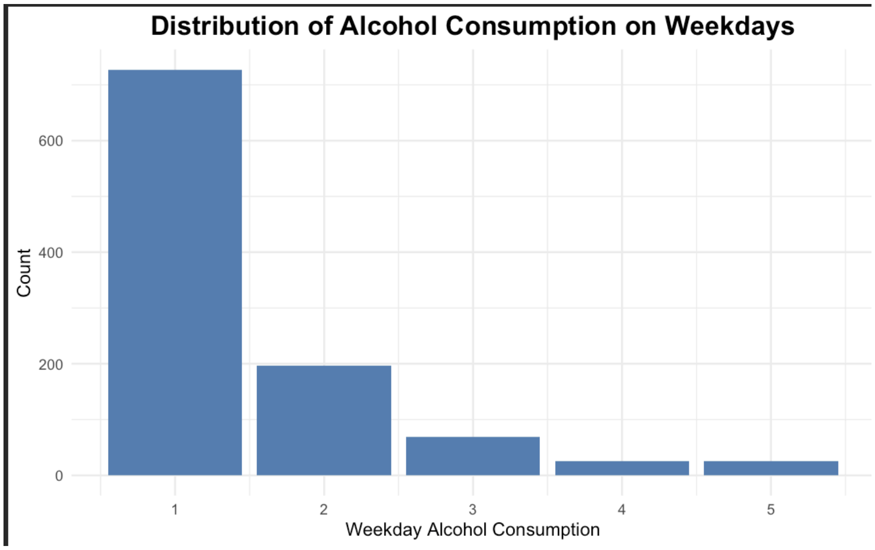
Firstly, the age distribution of students is heavily concentrated in the 15 to 18 age range, which accounts for more than 90% of the sample population. The highest frequency is observed at age 16, with over 250 students represented, while participation decreases significantly for ages 19 and older, with fewer than 20 students in the 19–22 age group. Overall, the graph is skewed right. This graph makes sense because it matches the typical age demographic of secondary school students.

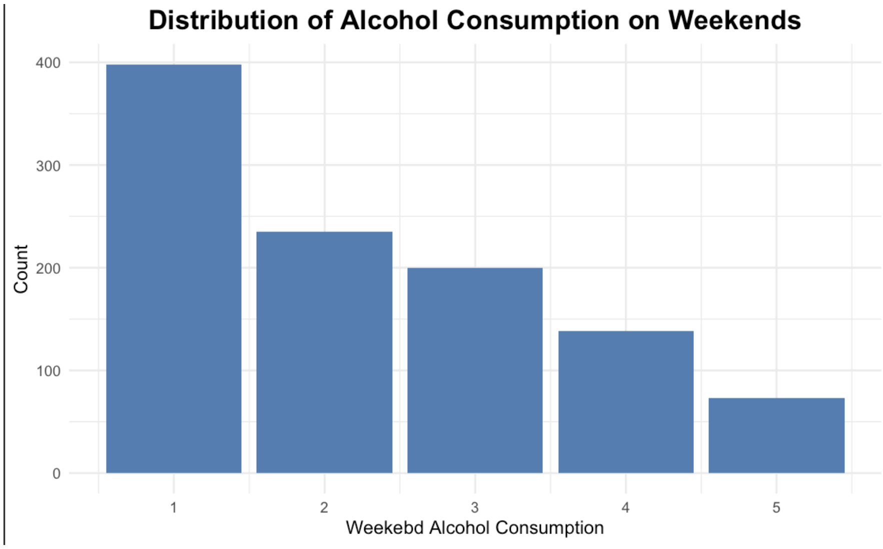


Next, the gender distribution of the sample was observed. The gender distribution graph indicates a slight imbalance, with most of the sample consisting of female students (n ≈ 600) and 47.3% male students (n ≈ 450).



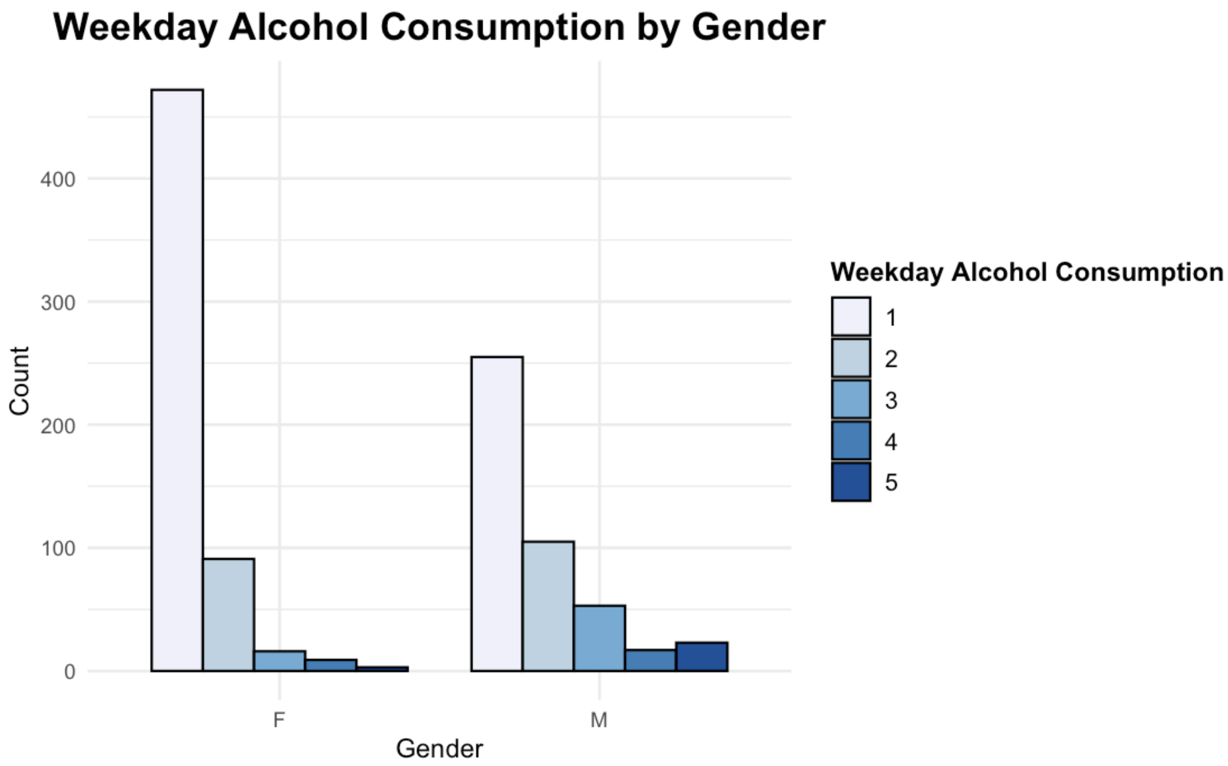
After this, the dependent variables were analyzed. After comparing the weekday and weekend Alcohol Distribution, we can see some obvious similarities and differences. For example, the weekend alcohol consumption graph (*Walc*) shows a much more balanced distribution compared to *Dalc*. Also, While category 1 ("very low") still includes a significant portion of students, categories 3 ("moderate"), 4 ("high"), and 5 ("very high") are more frequently represented. This shift indicates that students are more likely to consume alcohol on weekends, possibly due to fewer school-related obligations, greater social freedom, and increased opportunities for parties or gatherings. However, for both graphs, more of the sample reported in category 1 compared to all other categories.

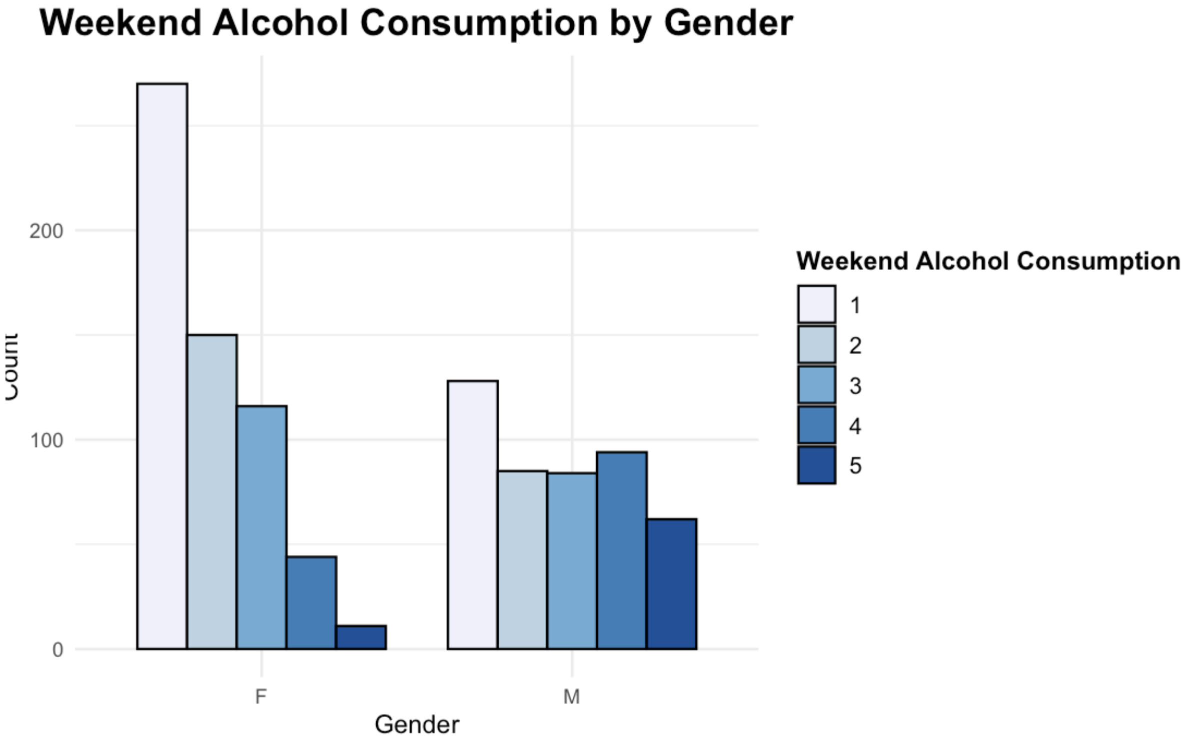




The bar graph depicting weekday alcohol consumption (Dalc) by gender illustrates a pronounced skew toward lower consumption levels for both males and females. Among female students, the majority (over 400) reported very low alcohol consumption (category 1), while progressively fewer students are observed in higher categories (2–5). Male students follow a similar trend, but the distribution is more dispersed, with fewer males in category 1 and a relatively higher number in the moderate-to-high consumption categories (3–5).

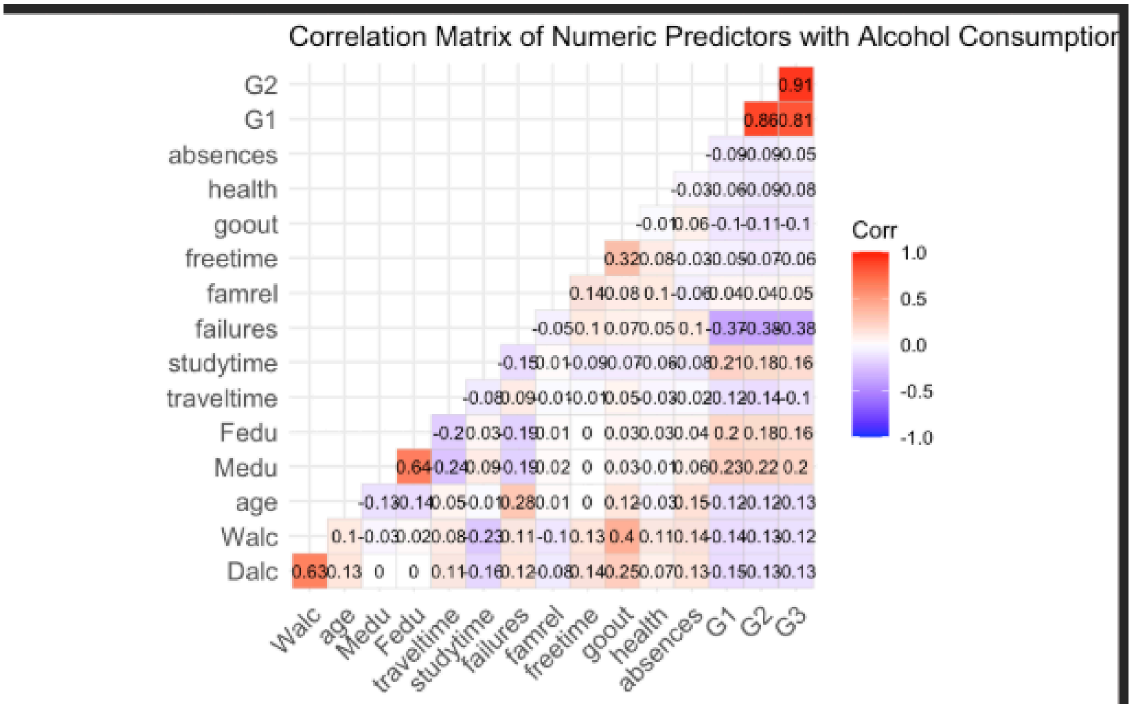
The weekend alcohol consumption (*Walc*) graph shows a shift toward higher levels of drinking compared to weekdays, with broader dispersion across all categories for both genders. Female students remain concentrated in the lower categories, with over 200 reporting very low consumption (category 1). However, more females appear in categories 3 and 4 compared to weekday drinking. Male students display a markedly different pattern, with fewer in category 1 and a much larger presence in the higher consumption categories (4 and 5).


Weekday Alcohol Consumption by Gender
Count
0
100
200
300
400
F
M
Gender
Weekday Alcohol Consumption
1
2
3
4
5


Weekend Alcohol Consumption by Gender
Count
0
100
200
F
M
Gender
Weekend Alcohol Consumption
1
2
3
4
5

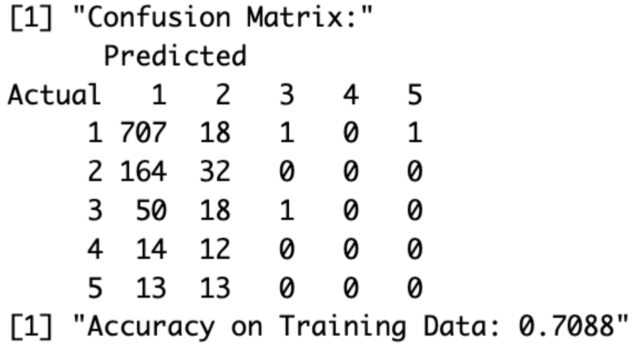
Next, the correlation matrix highlights relationships between numeric predictors and alcohol consumption (Dalc for weekdays and Walc for weekends). The strongest correlation observed is between Walc and Dalc (r = 0.63), indicating that students who drink heavily on weekends are likely to drink on weekdays as well. Academic performance, represented by grades (G1, G2, G3), shows a negative association with alcohol consumption. Final grades (G3) have a weak negative correlation with Dalc (r = -0.13) and Walc (r = -0.12), suggesting that higher alcohol consumption is associated with poorer academic outcomes.

Social variables also exhibit notable associations. The variable goout, representing the frequency of social outings, is positively correlated with Dalc (r = 0.32) and Walc (r = 0.35), indicating that students with more frequent social activities are more likely to consume alcohol. Maternal education (Medu) demonstrates a small negative correlation with both Dalc (r = -0.10) and Walc (r = -0.09), implying that higher maternal education may act as a protective factor against alcohol use. Lastly, school absences are positively correlated with Dalc (r = 0.07) and Walc (r = 0.09), suggesting a potential link between alcohol consumption and reduced school attendance.



5. Model Selection

To examine the relationship between alcoholic consumption and predictor variables, we initially employed an ordinal regression model, as the outcome variable *Dalc* (weekday alcohol consumption), is ordinal in nature with five levels. These levels range from 1 (low alcohol consumption) to 5 (high alcohol consumption). This ordinal regression model used *Dalc* as the dependent variable, and included Demographic, Performance, Social, and Support variables as predictors.



```
[1] "Confusion Matrix:"
      Predicted
Actual   1   2   3   4   5
     1 707  18   1   0   1
     2 164  32   0   0   0
     3  50  18   1   0   0
     4  14  12   0   0   0
     5  13  13   0   0   0
[1] "Accuracy on Training Data: 0.7088"
```

Following the implementation of the ordinal regression model, a confusion matrix was generated to evaluate the model's predictive performance. The matrix revealed a significant class imbalance, as the model predominantly predicted class 1, correctly classifying 727 instances out of the 1,044 total observations. In contrast, there were very few correct predictions for classes 2, 3, 4, and 5, and these predictions were notably less accurate. For instance, out of 164 actual instances of class 2, only 32 were correctly predicted. Classes 3, 4, and 5 were rarely predicted, indicating that the model struggled to accurately classify observations with higher levels of alcohol consumption, While the model achieved a high overall accuracy of 70.88%, this metric is heavily skewed by the dominance of class 1 predictions, masking its poor performance for the remaining classes.

The poor classification performance for higher alcohol consumption levels suggests that the model is biased due to the class imbalance in the dataset, with the majority of observations falling into class 1. To address this, we decided to undergo logistic regression instead and redefined the outcome variable into two classes. In this model, class 1 was retained to represent "low alcohol consumption," while classes 2, 3, 4, and 5 were combined into a single category representing "high alcohol consumption." This binary approach reduces the complexity of predicting multiple classes, thereby mitigating the effects of the class imbalance and enabling a clearer distinction between low and high alcohol consumption.

6. Variable Subset Selection

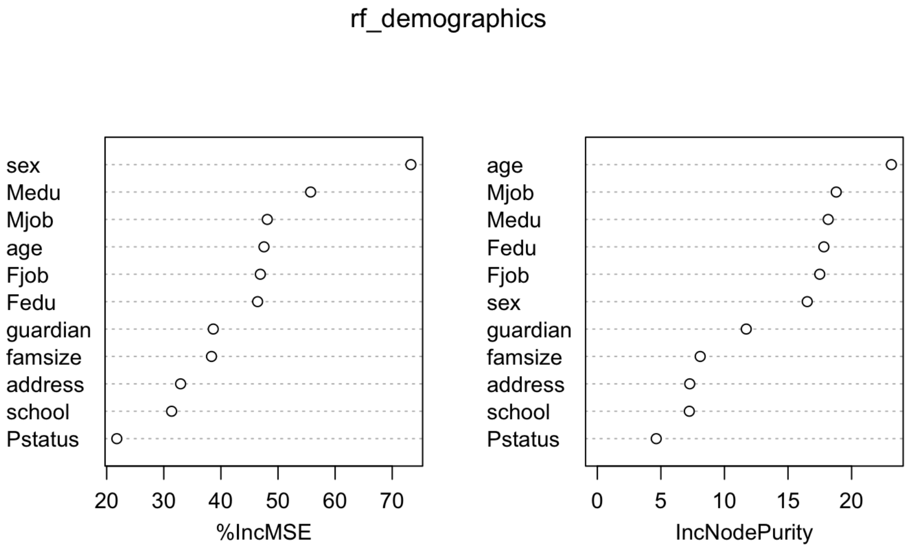
The next step in the process was to select which subset of predictors to include in the final model. The objective was to identify a subset of predictors that accurately represented the variable block it belonged to while balancing model simplicity and accuracy. Two variable selection methods were employed to fit a total of three models, which were then compared to determine the optimal variable subset.

The first model was created using random forest variable importance plots to identify the most important variables, and the model was then fit using only those predictors. Random forest variable importance plots were generated for each variable block: Demographics, Performance, Social, and Support.

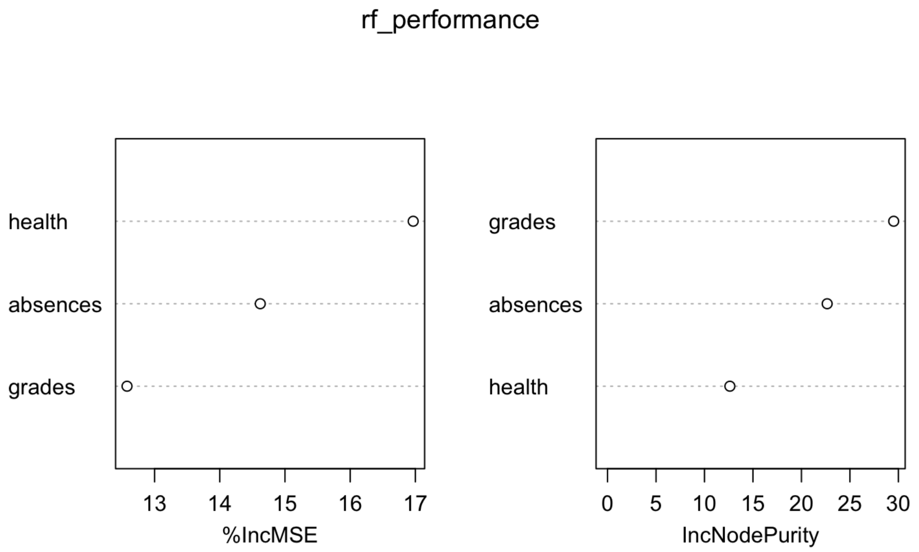
The second model was fit using stepwise backward regression, with the Akaike Information Criterion (AIC) as the selection criterion. Stepwise backward regression begins with the full model and iteratively removes the least important variable (based on its contribution to Mean Squared Error) until no further improvement in the model's AIC value can be achieved.

The third model was fit using the predictors that overlapped between Models 1 and 2, with the goal of reducing the number of predictors while maintaining or improving accuracy.

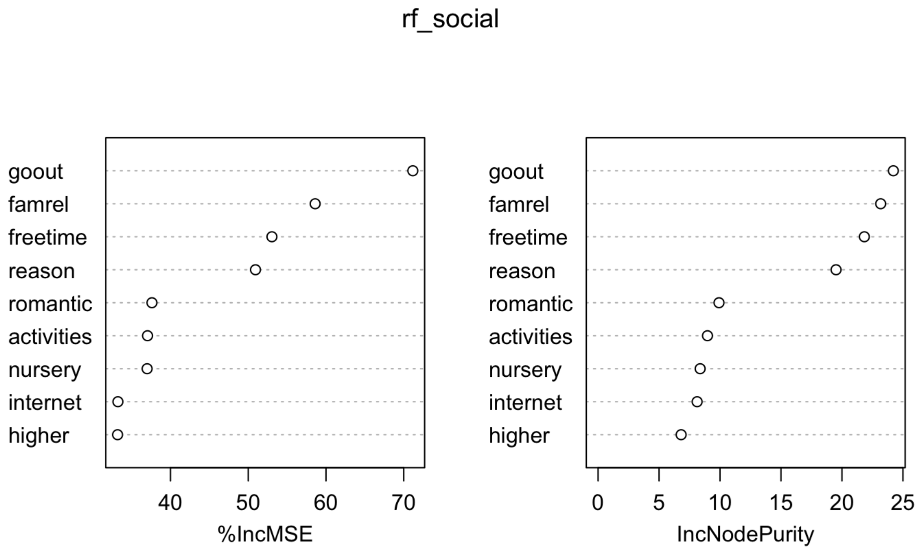
i. Model 1: Random Forest Variable Importance



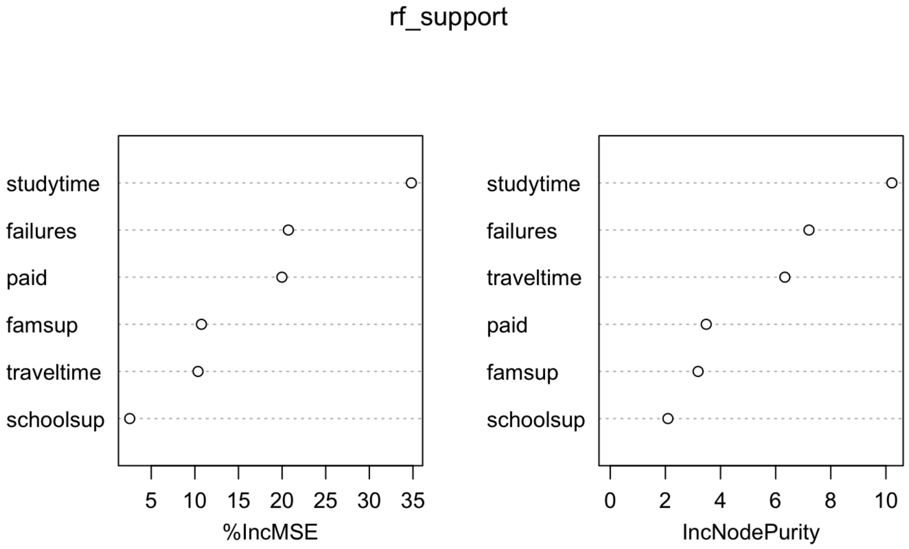
Predictors were selected based on their importance, measured by their contribution to the increase in Gini impurity and their percent increase in Mean Squared Error (MSE). From the "Demographics" variable block, the selected predictors were age, Mjob, Medu, Fedu, Fjob, sex, and guardian.



The predictors selected from the Performance variable block were grades and absences. While grades showed a high decrease in Gini impurity but a low percent increase in MSE, health exhibited a low decrease in Gini impurity but a high percent increase in MSE. To determine whether grades, health, or both should be included as predictors, the model was tested with each variable individually and then with both together. The model with the best accuracy included grades and absences but excluded health.



The predictors chosen from the Social variable block were go out, famrel, freetime, and reason.



The predictors chosen from the Support variable block were studytime, failures, and traveltime.

The final model fit using the predictors selected from the random forest importance plots is:

glm(Dalc ~ sex + address + famsize + Pstatus + Mjob + reason + guardian + studytime + paid + activities + nursery + famrel + freetime + goout + health + absences + grades, data = data2, , family = "binomial")

ii. Model 2: Stepwise Backward AIC

The second method used to determine the variable subset was Stepwise Backward AIC. This approach began by fitting the model with all the predictors, after which the step() function in R was used to perform stepwise regression. The function iteratively removed the least important predictors based on AIC, refining the model until no further improvement in AIC could be achieved.

The final model chosen's formula is

glm(Dalc ~ sex + address + famsize + Pstatus + Mjob + reason + guardian + studytime + paid + activities + nursery + famrel + freetime + goout + health + absences + grades)

The results of the final step of the backwise regression were analyzed to ensure that removing no other variable was necessary.

Variable Removed	**Df**	**Deviance**	**AIC**
<none>		1024.6	1072.6
- nursery	1	1027.0	1073.0

- freetime	1	1027.3	1073.3
- Pstatus	1	1027.3	1073.3
- studytime	1	1027.3	1073.3
- paid	1	1027.8	1073.8
- health	1	1028.8	1074.8
- grades	1	1028.8	1074.8
- activities	1	1030.1	1076.1
- absences	1	1032.8	1078.8
- address	1	1033.8	1079.8
- Mjob	4	1042.0	1082.0
- guardian	2	1040.0	1084.0
- famsize	1	1039.2	1085.2
- reason	3	1044.0	1086.0
- famrel	1	1057.1	1103.1
- sex	1	1064.2	1110.2
- goout	1	1074.8	1120.8

iii. Model 3: Combined Model

The final model was fit using the predictors selected through both the random forest variable importance plot process and the stepwise regression process. The objective of this approach was to reduce the number of predictors while assessing whether accuracy would decrease, remain consistent, or ideally, improve. The formula for the combined model is:

glm(Dalc ~ sex + Mjob + reason + guardian + studytime + famrel + freetime + goout + absences + grades, data = data2, family = "binomial")

iv. Model Comparisons

To determine which subset of variables should be used in the final model, the three candidate models were compared using several metrics: the number of predictors, residual deviance, AIC, confusion matrix, accuracy, and Leave-One-Out Cross-Validation (LOOCV) error rate. The number of predictors, residual deviance, and AIC were obtained by analyzing the summary of the logistic regression models in R. The confusion matrix for each model was created by predicting the responses and comparing the predictions with the actual responses. Accuracy was calculated as the ratio of correct predictions to the total number of predictions.

The LOOCV error rate was determined by performing Leave-One-Out Cross-Validation on each model. LOOCV involves removing one observation from the dataset, fitting the model on the remaining observations, predicting the response for the removed observation, and calculating the Mean Squared Error (MSE) between the predicted and actual responses. This process is repeated for all observations, and the resulting MSEs are averaged to estimate the overall MSE for the model. The purpose of running LOOCV is to estimate the testing MSE, which is particularly important since the dataset was not split into training and testing subsets.

The results of the Model Comparison were:

<table>
<tr><th></th><th>M1: Random Forest Variable Importance</th><th>M2: Stepwise Backward Elimination</th><th>M3: Combined</th></tr>
<tr><td><b># of predictors</b></td><td>15</td><td>17</td><td>10</td></tr>
<tr><td><b>Residual Deviance</b></td><td>1056.1</td><td>1024.6</td><td>1060.9</td></tr>
<tr><td><b>AIC</b></td><td>1108.1</td><td>1072.6</td><td>1094.9</td></tr>
<tr><td><b>Confusion Matrix</b></td>
<td>Predicted<table><tr><td rowspan="3">Actual</td><td></td><td><b>0</b></td><td><b>1</b></td></tr><tr><td><b>0</b></td><td>661</td><td>66</td></tr><tr><td><b>1</b></td><td>188</td><td>129</td></tr></table></td>
<td>Predicted<table><tr><td rowspan="3">Actual</td><td></td><td><b>0</b></td><td><b>1</b></td></tr><tr><td><b>0</b></td><td>654</td><td>73</td></tr><tr><td><b>1</b></td><td>176</td><td>141</td></tr></table></td>
<td>Predicted<table><tr><td rowspan="3">Actual</td><td></td><td><b>0</b></td><td><b>1</b></td></tr><tr><td><b>0</b></td><td>665</td><td>62</td></tr><tr><td><b>1</b></td><td>185</td><td>132</td></tr></table></td></tr>
<tr><td><b>Accuracy</b></td><td>0.7567</td><td>0.7615</td><td>0.7634</td></tr>
<tr><td><b>LOOCV Error Rate</b></td><td>0.1764</td><td>0.1701</td><td>0.1739</td></tr>
</table>

Model 2, the Stepwise Backward Elimination model, had the lowest residual deviance and AIC, likely because it included the highest number of predictors. Both residual deviance and AIC are known to favor models with more predictors as they do not impose strong penalties for model complexity.

Model 3, the Combined model, achieved the highest accuracy rate, but the difference between the three models was minimal—within one percentage point. Similarly, Model 2 had the lowest LOOCV error rate, but all three models had comparable error rates within 0.6 percentage points of each other.

Ultimately, Model 3 was selected as the optimal final model. While its residual deviance was slightly higher than the other models (less than 4 points higher than Model 1), this was offset by the fact that it used significantly fewer predictors—5 fewer than Model 1 and 7 fewer than Model 2. The AIC and residual deviance of Model 3 were comparable to those of the other models, and its LOOCV error rate was also on par.

To summarize, the simplicity of Model 3, combined with its highest accuracy and minimal trade-offs in AIC, residual deviance, and LOOCV error rate, made it the most appealing choice for the final model.

v. Selection of Variable Blocks

Based on the results of the variable subset selection within variable blocks, the final model candidate is:

> *glm(Dalc ~ sex + Mjob + reason + guardian + studytime + famrel + freetime + goout + absences + grades, data = data2, family = "binomial").*

In this model, the following predictors are included in each variable block:

- Demographics: sex, Mjob, guardian
- Personal and Academic Performance: grades, absences
- Social and Relationships: reason, go out, family relationships, free time
- Support: studytime

The last step of the variable subset selection was to determine if each variable block should in fact be included in the final model, or if one or more of the variable blocks was obsolete. To do this, a model was fitted on every combination of variable blocks possible (2^4 = 16 combinations), excluding the model with no variable blocks included. The results of these comparisons were analyzed to determine the contribution of each block to the overall model performance.

Combination	Residual Deviance	AIC
demographics, performance, social, support	1060.89077503054	1094.89077503054
performance, social, support	1140.1620200723	1160.1620200723
demographics, social, support	1074.86897266965	1104.86897266965
social, support	1154.72187780524	1170.72187780524
demographics, performance, support	1155.20668464653	1177.20668464653

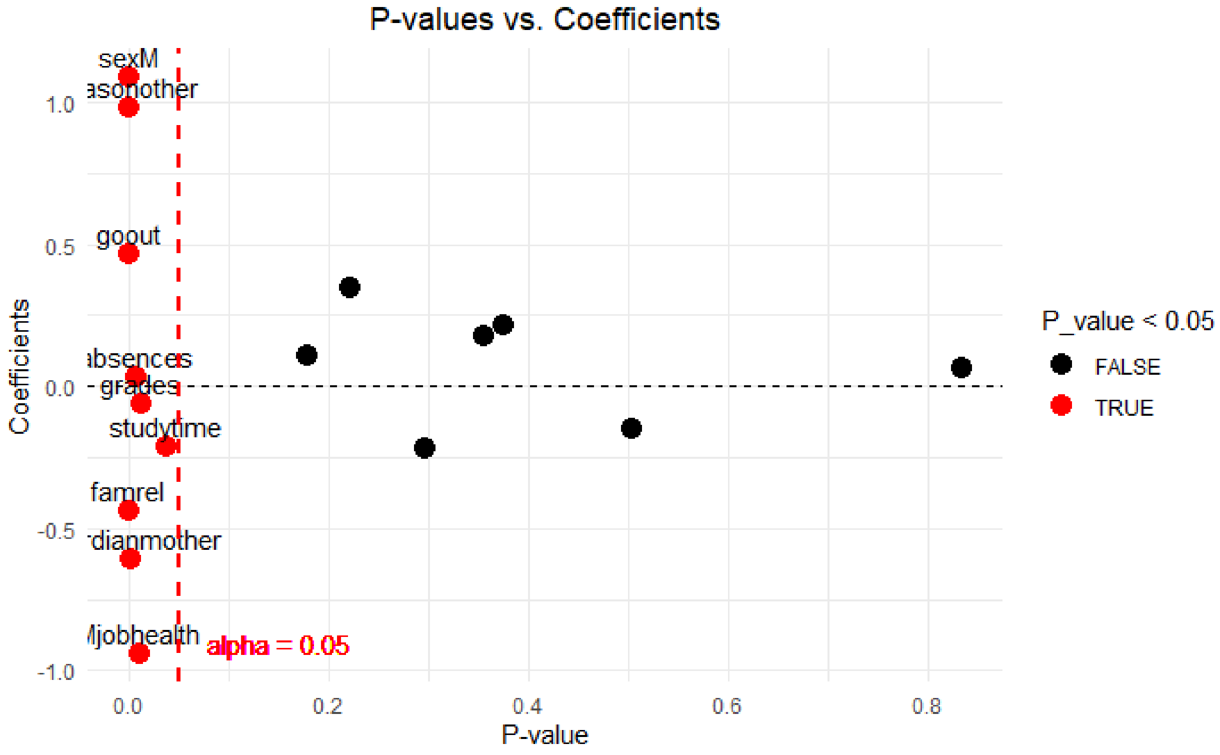
performance, support	1228.25392830948	1236.25392830948
demographics, support	1180.93713335481	1198.93713335481
support	1251.55268317422	1255.55268317422
demographics, performance, social	1065.30457857109	1097.30457857109
performance, social	1152.54793691434	1170.54793691434
demographics, social	1082.40281990601	1110.40281990601
social	1172.09512530233	1186.09512530233
demographics, performance	1163.39579547892	1183.39579547892
performance	1249.01710482481	1255.01710482481
demographics	1195.75197844633	1211.75197844633

The model with the highest AIC and residual deviance was the one that included all four variable blocks. As a result, the final model candidate remained unchanged.

7. Results

Given the discussed model and feature selection, a logistic regression model with 10 predictors was elected to be used.

i. Coefficients



Wald tests were conducted in R to assess coefficient significance with a p-value < 0.05. Given this, 9 predictors were determined significant; of these, 4 were categorical levels (sexM, reasonother, guardianmother, Mjobhealth) and 5 were numerical predictors (goout, absences, grades, studytimes, famrel).

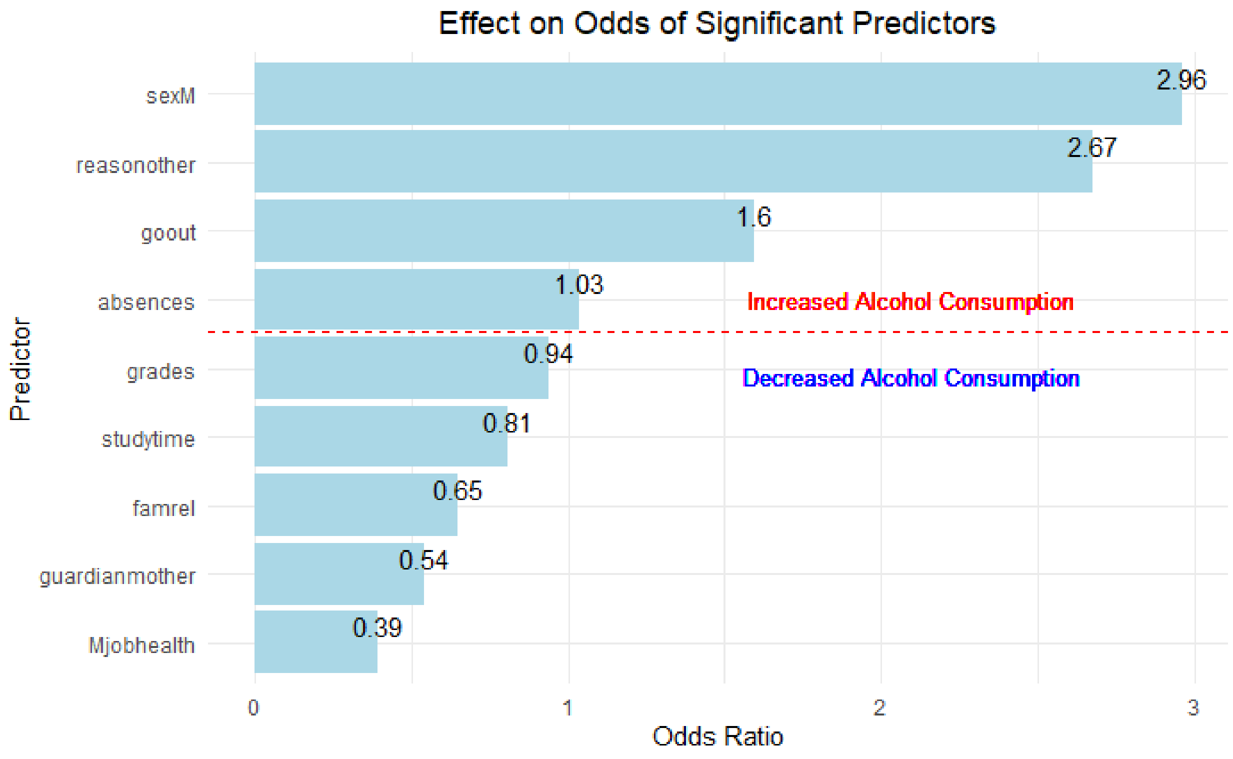
A logistic regression model is structured as follows:

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Therefore, the effect of each predictor on the odds of the response is evaluated by taking the exponential function of that predictor's coefficient.

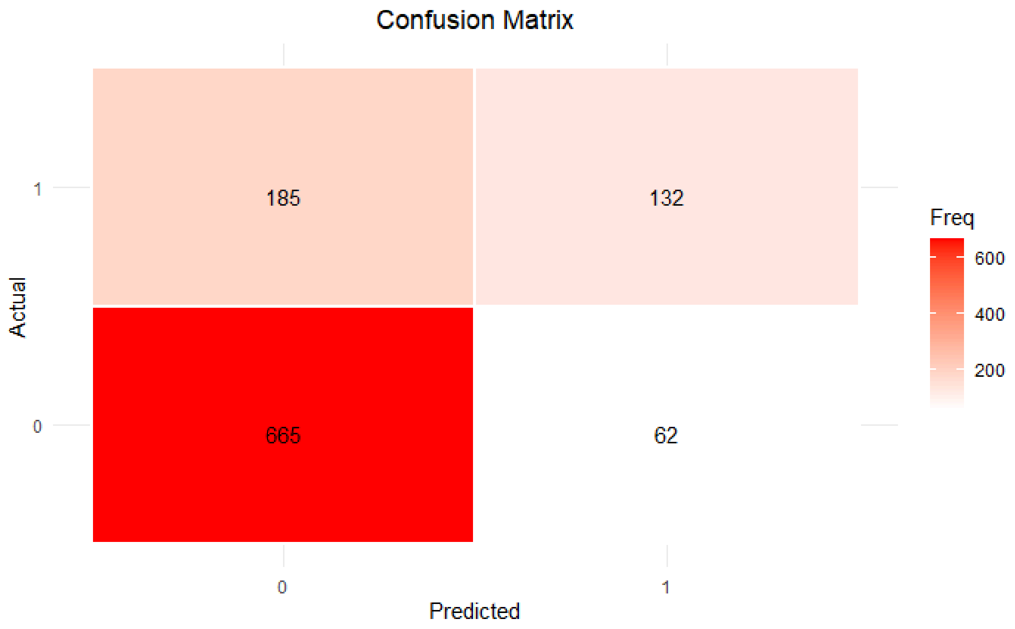
$$\text{Effect on } \frac{p}{1-p} \text{ by } \beta_i = \exp(\beta_i)$$

This was applied to the coefficients of the 9 significant predictors in this model to interpret predictor contribution to the odds of increased student alcohol consumption. The odds ratio explains the proportion by which the presence of a categorical level or the increase of a numerical predictor's value by one point changes the odds of a student partaking in a higher level of alcohol consumption from the base case.



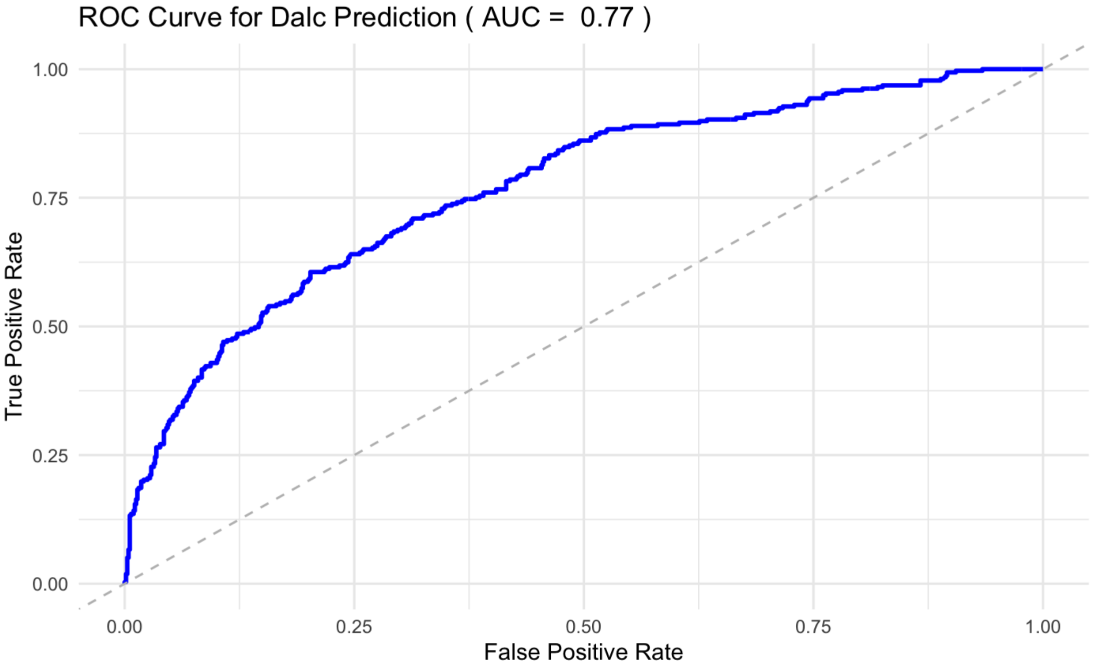
ii. Accuracy

Using a seed of 1234, fitted values for the response were predicted with the model given the training dataset. The resulting confusion matrix was as follows:



Thus, the training accuracy of this model is 0.7634; its true positive (1) rate is 0.6804, the false positive rate is 0.2176, the true negative (0) rate is 0.7823, and the false negative rate is 0.3195.

Using these figures for true positive and false positive rate derived from the confusion matrix, an ROC curve was fitted.

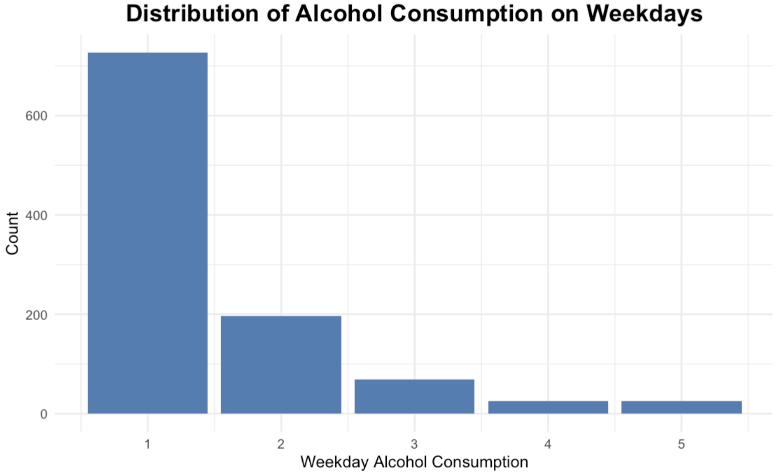


The ROC curve yielded an area-under-curve (AUC) of 0.7686. This assessment of accuracy is comparable and consistent with the accuracy we derived from the confusion matrix.

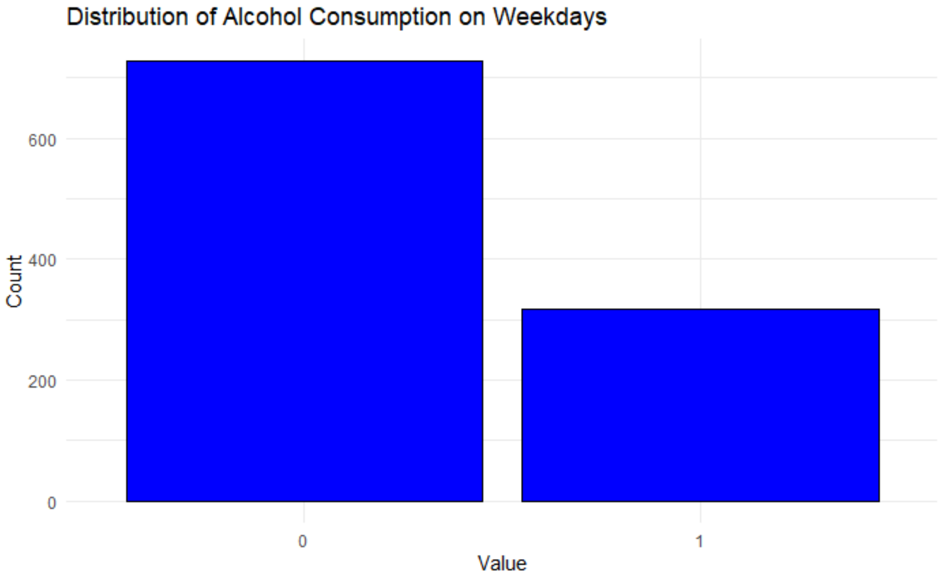
8. Conclusion

i. Limitations

The ordinal response variable Dalc in the original dataset was heavily left skewed, which resulted in the initial ordinal regression overpredicting ‘1’ as a response.



Collapsing the classes of this ordinal variable to reduce it into a binary distinction that can be predicted with logistic regression was an effective approach to this problem and helped fit a more predictively useful model. However, the natural distribution of the data resulted in the binary variable still being far skewed towards lower alcohol consumption.



This skew is consistent with the true positive rate (0.6804) and true negative rate (0.7823) derived from the confusion matrix. A true positive rate lower than the true negative rate indicates that the model is more accurate at predicting cases of low alcohol consumption than high alcohol consumption. This and the skew of the data indicate that the model is underfitted to the class of high alcohol consumption, inflating the count of falsely predicted low-alcohol-consumption cases.

As a result, this model will predict few false positives for high alcohol consumption, but in turn will fail to identify some students who do consume a high level of alcohol.

ii. Coefficients in Context

sexM was the most significant and effectual predictor in indicating an increased odds of higher alcohol consumption. This shows that male students are an extremely high-risk group towards high alcohol consumption. reasonother is a predictor that also had a very large effect on this odds, and indicates students who answered “other” when surveyed why they chose their school. This predictor seems the most arbitrary of the significant predictors in the context of importance to the response variable, and it could be beneficial in future study to investigate more closely into this relationship.

Mjobhealth, guardianmother, and famrel had the most effect in reducing the odds of high alcohol consumption. These predictors respectively represent students with mothers who work in health, students whose primary guardian is their mother, and the quality of a student’s relationships with family. This indicates that a student’s mother plays an especially significant role in determining their level of alcohol consumption, as does their overall family environment and involvement, which is consistent with the results of Posavec’s study (2022).

goout and absences significantly increased the odds of high alcohol consumption, while studytime and grades significantly decreased that odds. These predictors tell us how often a student goes out with friends, how many absences they have, how much they study, and their grades, which all concern their status, attitudes, and performance in school, as well as their social life. While all of these predictors are significant, their effects on odds are all small, especially in comparison to the other significant predictors. This is consistent with the findings of Balsa et al. (2012).

References

UCLA Institute for Digital Research and Education. "How Do I Interpret Odds Ratios in Logistic Regression?" *UCLA Statistics Online Resource Center*, https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/. Accessed 9 Dec. 2024.

Posavec, Jelena Šarić, et al. "Use of Diagnostic Tests in Primary Care: A Case Study." *National Institutes of Health*, https://pmc.ncbi.nlm.nih.gov/articles/PMC10006657/. Accessed 9 Dec. 2024.

Esser, Marissa B., et al. "MMWR Weekly Report." *Centers for Disease Control and Prevention*, https://www.cdc.gov/mmwr/volumes/66/wr/mm6618a4.htm. Accessed 9 Dec. 2024.

Balsa, Ana I, et al. "Characterization of a Novel Therapeutic Target in Cystic Fibrosis." *National Institutes of Health*, https://pmc.ncbi.nlm.nih.gov/articles/PMC3026599/. Accessed 9 Dec. 2024.