**1. Dataset used and the problem addressed in this analysis.**  
  
The dataset we chose is about the student alcohol consumption. The dataset was shared by UCI Machine Learning on Kaggle.com and it is from a survey conducted in Europe, about student background and alcohol consumption.

The main point we want to learn from this data is:

*How are the students’ personal background (family size, age, gender, etc.), if any, correlated to their alcohol consumption*

**2. Exploratory data analysis**  
  
The dimension of the dataset was 395 records and 33 variables pertaining to different personal backgrounds such as age, gender, family size, and so on about each student as well as the rate of alcohol consumption.

Upon examining the structure of the data using structure function, we found out that 16 of the 33 variables were integer variables.

Using the summary function, we gained the five number summary for continuous variables; counts of binary responses; counts for the Likert scale; and nominal variables.

- As for the continuous variable summary, there was an outlier for the ‘absences’ variable.

- For binary responses, the NO was nearly 8 times more than YES for weekday alcohol consumption whereas NO response was only 1.5 times more than YES for the weekend alcohol consumption

- Some responses in the form of a rating scale

We noticed that the response variables were recorded as integers and because of the categorical nature of our response variable, we converted the continuous variables to categorical variables using as.factor() function.

Also, our response variable, the alcohol consumption rate, was separated into two – weekday alcohol consumption and weekend alcohol consumption with a rating scale of 1 to 5.

- So we had to convert our response variable into a binomial variable.

- We set a consumption rate less than or equal to 2 as “NO” and greater than 2 as “YES” for alcohol consumption.

- After that, we converted these two variables as categorical ones like before.

For simpler analysis, we converted these responses to a binary type, where we consider the consumption rate less than or equal to 2 as ‘NO’ and greater than to as ‘YES’ for alcohol consumption. At the same time, we made two variable contrasts or dummy variables where 0 takes ‘NO’ and 1 represents ‘YES’.

**3. Reducing variables**  
   
Our primary assumption of possible predictors that could have some correlation with a higher alcohol consumption rate was as follows:

Sex: the student’s gender

Pstatus: the parent’s cohabitation status – whether the parents are living together or apart

Romantic: the student’s relationship status

Absences: the number of school absences

Failures: the number of past class failures

Famrel: the quality of a family relationship

However, to get a more meaningful set of predictors, we decided to use the stepwise regression.

We first created two variables

- One, a null variable containing only the intercept.

- And a full variable containing all the variables from the dataset and regress them in a glm function with the family option set to “binomial”

Then we set up a variable for the step function where we set the vector, the scope, and the direction to “both”.

- We repeated this process for the 2nd response variable which is ‘weekend alcohol consumption’ or Walc.

Since we have two response variables, we repeated this process twice. For the first response variable, weekday alcohol consumption, the starting AIC was 278.04. By the end of the regression, the AIC was reduced down to 226.76 with the following predictors:

- sex; goout; school; absences; traveltime; activities; higher; reason; famsize; nersery

For the second response variable, weekend alcohol consumption, the starting AIC was 534.48 and the final AIC was reduced down to 437.07 with the following variables:

- goout; sex; fjob; absences; famrel; paid; traveltime; address; famsize; nursery; activities

In result, only 50% of the predictors we initially assumed was also included in the best subset selection - and they were, sex, absences, a family relationship; but parent’s cohabitation status, romantic, and failure was not included in the result.

We had the largest odds ratio for sexM predictor where, holding all other predictors fixed, the one-unit increase in SexM, increased the odds of weekend alcohol consumption by a factor of 8. In other words, we expect to see about 8 times more in the odds of weekend alcohol consumption for male students.

**4. Select appropriate data mining methods to develop models. Compare model performances. Implement multiple models until you reach a “final or optimal” model. Describe the model development and selection process.**

Since our response variables are categorical in nature, we selected logistic regression and decision tree models as our data mining methods. Since our dataset is small, we needed to employ a resampling method for our dataset.

We initially used only the bootstrap as our resampling method but we noticed that our dataset was highly unbalanced. We use the ROSE package to resolve this issue. By applying the ROSE package, we can balance the data. After the dataset is balanced, we apply it to the bootstrap to create balanced sample datasets.

After obtaining balanced sample training datasets, we use these datasets to fit our model.

***Logistic Regression***

For the logistic regression, we used 4 different resampled datasets to train our model.

We use the glm() function containing the predictors obtained from stepwise regression. We set data equal to a balanced dataset; set subset equal to a training dataset created from the balanced dataset; and family equal to binomial.

We use subset option in glm() to fit a regression using only the observations corresponding to the subset and set the family equal to binomial since we have a binary outcome.

After that, we use predict() function to evaluate the performance of the model using the test dataset. It calculates the predicted probabilities of alcohol consumption, and we set the type equal to “response” which will give us the predicted probability.

***Classification Decision Tree***

For the classification tree analysis, we use balanced dataset created from using ROSE and bootstrap as before.

We use the tree() function where we regress our response variable on predictors we obtained from stepwise regression. We set the subset equal to the balanced sample dataset.

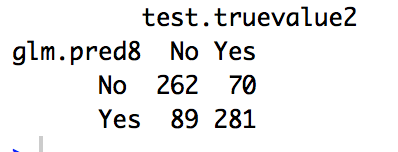
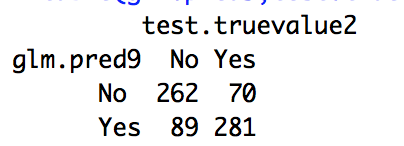
Then we performed cross-validation with K=10 to obtain the optimal tree size.

* **Finding result & Interpretation** 
  + Showing the decision trees for both Daily/Weekly Alcohol Consumption(MSE)
  + R code consist of training and testing dataset, applied library *rpart* for displaying decision trees
    - Features and each function
      * **rpart.plot(tree.model2, type=4,extra=101,nn = TRUE,branch=1,varlen=0,yesno=2)**
* Brief Summary:
  + Consuming alcohol does harm to students physically and mentally.
  + Discussing the crucial factors as a result of the family’s background could be more likely to cause students to consume alcohol.

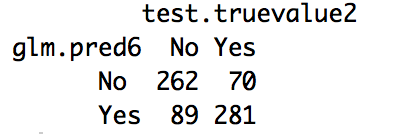
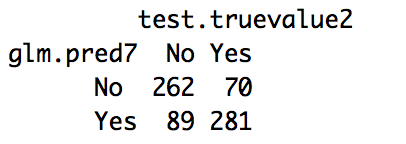
**5. Adopt approximate methods and measures for model evaluation. Use figures or tables to show model performance.**   
  
We used bootstrap to check the accuracy for the logistic models and the decision tree due to the limited sample size. The high imbalance in the response variable called for the use of a balanced training dataset to fit the model. We have produced several confusion matrices with balanced training datasets below.

Figure 1-4 is the confusion matrices produced by using the balanced training datasets with the model for “Weekday alcohol consumption”, and the accuracy is 0.79

**Do the same for decision tree model!!!!!!!**

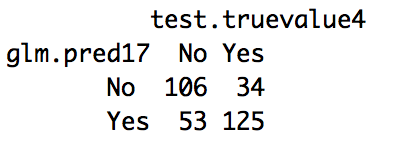
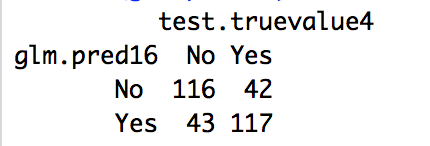


(Figure 1) (Figure 2)

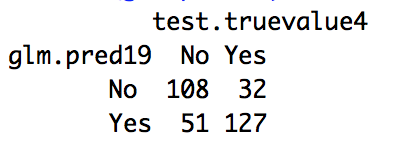
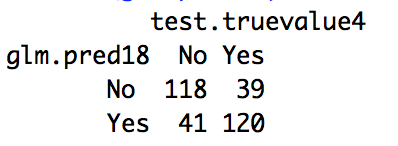


(Figure 3) (Figure 4)

Figure 5-8 is the confusion matrices produced by using the balanced training datasets with model for “Weekend alcohol consumption”, and the accuracy is 0.74



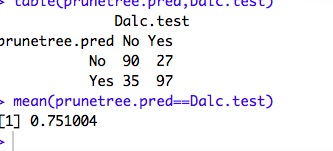
(Figure 5) (Figure 6)



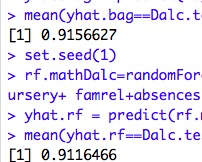
(Figure 7) (Figure 8)

**Weekday**

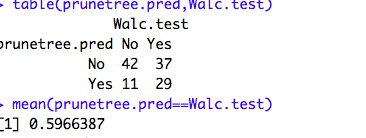
**Pruned Tree model**

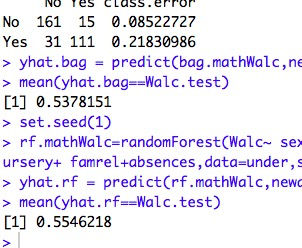


Bagging & RandomForrest

****

**Weekend**

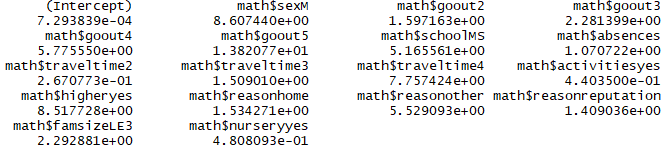
****

****

Need to fix unbalanced issue first

**6. State your final model clearly and show your findings. Use figures, tables or formatted outputs to support your arguments.**   
**Compare logistic regression and decision to see which one is better, use the accuracy list as the figure!! And just simply describe the model, if logistics regression can mention odds ratio(Soo knows), if decision tree maybe can describe the node or something(ask minty)**

**7. Make conclusions following logically from results and findings. Discuss its implications to the target audience.**   
**State which model is better**

****

Based on our models, besides some uncontrollable factors like gender, family-related factors like family size, and the quality of family relationship were important predictors of higher rate of alcohol consumption.

Likewise, the number of school absences is a good tracking metric for students’ alcohol consumption where, the more a student skips the class, the more likely it was for the students to consume alcohol regularly.

? Above conclusion based on predictors from stepwise regression?

? Do you think we can use the factors listed above (odds ratio)?