**Section 1 – Introduction:**

Multiple variables can affect a student’s grades. Under specific circumstances, we can possibly predict secondary students’ grades. More specifically, is alcohol consumption amongst students a significant factor in students’ grades? Our research question is “how can student alcohol consumption, combined with other factors, affect student grades?”

The “Student Alcohol Consumption” datasets being observed comes from the Kaggle database. This data came from a survey of students in math and Portuguese language courses in two secondary schools. The two schools, Gabriel Pereira and Mousinho da Silveira, are both located in Portugal. Variables being observed include sex, age, family size, parent’s cohabitation status, parent’s education and jobs, time spent studying, participation in extra-curricular activities, health, absences, grades, workday and weekend alcohol consumption. Grade variables are separated into three columns, G1, G2, and G3. These columns are the first period grade, the second period grade, and the final grade. All grades are numeric on a scale of 0 to 20.

**Section 3 – Data analysis plan:**

There are many independent variables in this study and some of them are age, sex, family size, mother and father’s job, parent contribution status, relationship status, and many more. The target variable is student’s grades. In this study we will analyze the mathematics class data set. We will be comparing the data from the mathematics class with factors that affect student grades. We will also be comparing the grades from different periods of time from the set, semester one, semester two, and overall performance.

There are many different types of data in the data set we chose including numeric, binary, and nominal data. For the binary data, I think that the decision tree and random forest method would be helpful in this study. In order to do this, we will have to make a binary classifier variable, splitting the overall grade into two categories, one being poor grades and the other being good grades. The decision tree would be effective because it is used for predicting two categorical variables and it takes a majority based on the splits. Random forest would help to compare the data from the decision tree. For the numeric data, I think the multiple regressions model would be effective. This allows room for many variables and we can check their significance using the backward elimination and looking at the p-value.

The results needed to support our hypothesized answer are the p values that show us which variables contribute most to getting the desired grade. This in return will help us predict the numeric grades given by the model. The classifiers will help predict the final grades of the students.