**PREDICT HOW ALCOHOL CONSUMPTION AFFECTS STUDENT’S AVERAGE GRADES**

**BY**

**AVINASH YAVVARI**

**KRISHNA CHAITANYA ASHISH VATTEM**

**RAMANI REDDY**

**RAMYA SRI SONAR**

**UNDER PROFESSOR**

**RAN JI**

**OR 568 : APPLIED PREDICTIVE ANALYTICS**

****

**GEORGE MASON UNIVERSITY**

**05/14/2020**

**ABSTRACT:**

Alcohol consumption is known to have caused predominant health issues in the United States. Drinking alcohol excessively has a great impact on the health and social behavior of students. Around 20% of the students suffer from alcohol use disorder and it was also seen that their academic performance was poor. In order to see how valid these facts are, we have taken the Student Alcohol consumption dataset and performed data analysis. This dataset contains information about students’ grades, weekly alcohol consumption, daily alcohol consumption and other features as well.

We will see from the analysis that alcohol is not the only reason that affects the grades of the students. We understood that alcohol will affect the grades of only some students, but it is not the same for all the students. We have also predicted the final grades using these features by using classification algorithms by classifying the final grades as good grades and bad grades. Analyzing which algorithm works best for our dataset and understanding how it performs well will also be seen. Exploratory analysis is done to find some patterns and gain new insights from the visualizations which will be explained as well.

**INTRODUCTION:**

In a survey conducted between the math and Portuguese language courses in secondary school, a lot of interesting things like social, gender and study information have been gathered along with the alcohol consumed by the students. The dataset has been obtained from UCI repository and we have two datasets- one for math course and the other for Portuguese course. The dataset contains 33 columns in both the datasets and the number of students in the math course are 395 and the students in the Portuguese class are 649.

The datasets will be combined further, to analyze how the alcohol consumption affects the grades of the students. When the datasets are merged, we find that 382 students belong to both the math and Portuguese courses. We will be predicting the final grade as well. Though the final grade is continuous, we will be doing binary classification by classifying the grades into two classes namely good grades and bad grades. We will explain the algorithms used, why they were used, how they performed and why we split the final grade into 2 classes in the coming sections.

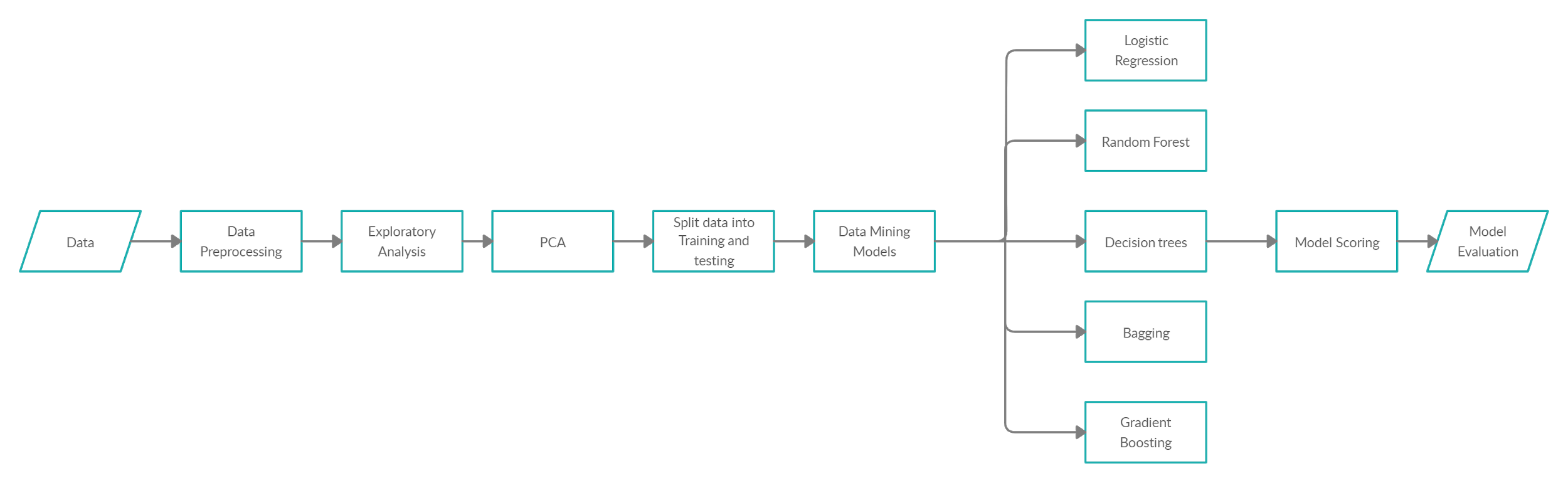
**OBJECTIVES:**

The main objective of this project is to see if alcohol affects the final grades of students based on the amount of alcohol students consume weekly or daily. The second objective is to see which other features will affect the final grades of students. The next objective is to correlate the multiple features available in the datasets to find the most significant features that will affect the final grades of the students. The next objective is to gain insights by doing exploratory data analysis and understand how the data is distributed. The last objective is to implement multiple machine learning algorithms to find which algorithm works best in predicting the final grades of the students. Meanwhile, we will also see the importance of first term grades and second term grades.

**DATASET:**

As mentioned earlier, the datasets have been taken from the UCI Machine Learning Repository and they are 2 datasets, one for the math course and the other for the Portuguese course. The dataset consists of 33 columns in both the datasets and they are 426 students in both the classes. In these 426 students 382 students are a part of both the classes after dropping the duplicate rows we will have 85 students left in both the classes. We have features such as age, weekday & weekend alcohol consumption, father’s & mother’s education, health status, number of times the student was absent for the class, family size of the student, time student allocated for studying and travelling, internet access at home and many more. Some of these features might depend on the other available features and some might be completely independent. The alcohol consumption which is the main feature has values ranging from 1 to 10 where 1 means very low consumption and 10 means very high consumption of alcohol. The data types of features range between strings, integers and Boolean.

**CONCEPTUAL DIAGRAM:**



**Figure 1. Conceptual Diagram**

**DATA PREPROCESSING:**

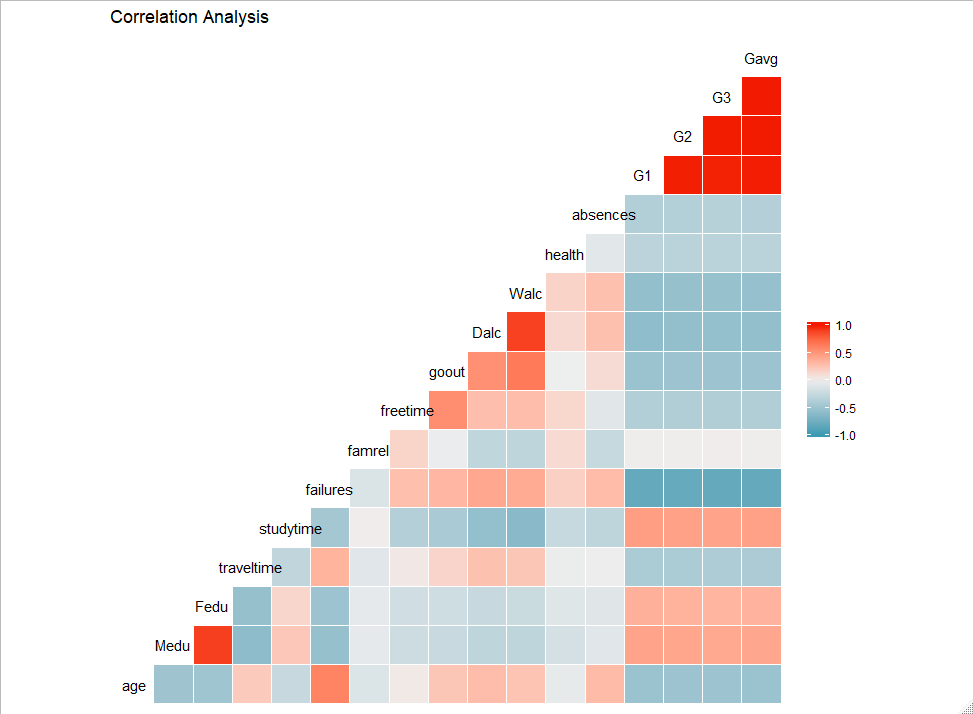
The first step of data preprocessing is to identify the missing values in the data and replace it with mean or else drop those records. As we do not have any missing data, we will check the different columns in the dataset and analyze them to know which ones can be useful in predicting the final grades and which ones can be dropped as some features are of no use to us. We have added a new column called Gavg that contains the average of first term grade, second term grade and the final grade to know how the distribution of grades. The features like address, sex and school name are not considered as they do not have much importance in predicting the final grade. We have also divided the alcohol consumption into 5 categories such as very low, low, medium, high and very high.

**EXPLORATORY ANALYSIS:**

We will be using ggplot for exploratory analysis and for visualizations, multiple features will be plotted against the grades to see how each feature affects the grades and we will also see if the most likely features we think will affect the grades are actually the ones which will affect the grades. Let us begin with the correlation analysis. We also see the weekday and weekend alcohol consumption plots to get a better idea of how alcohol affects the final grades.

**Correlation Analysis:**

A correlation plot for checking the correlation between the multiple features of the dataset is shown below. Correlation analysis is a statistical method used to evaluate the strength of relationship between two features or variables. A strong correlation means that two or more variables are most likely related to each other and a change in one variable will result in change of another variable , while a weak correlation means that the variables are not that related to each other.

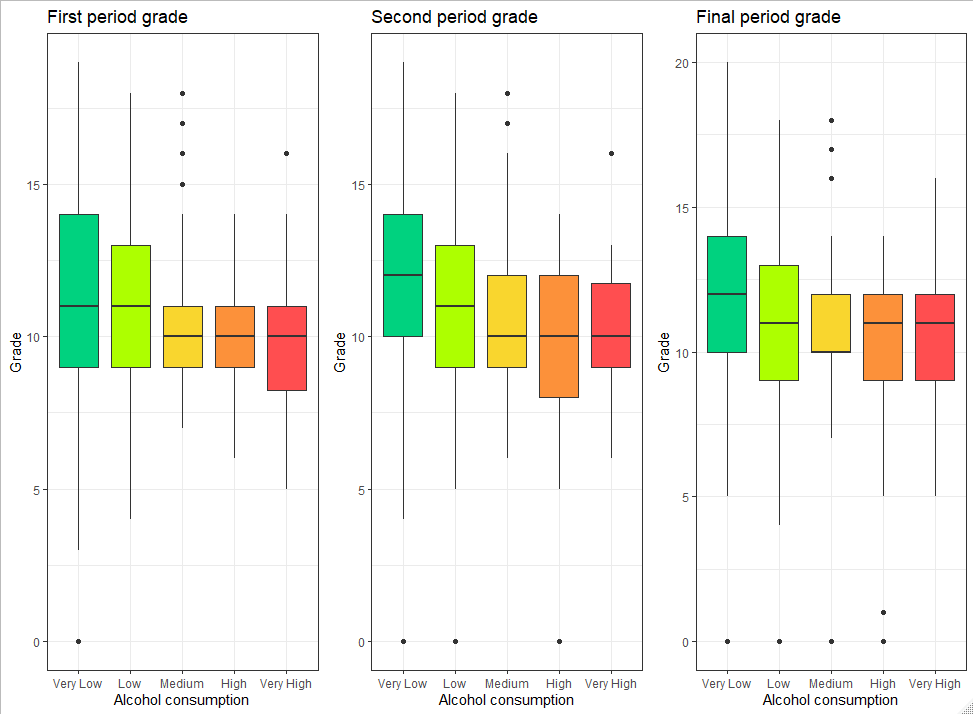


**Figure 2. Correlation Plot**

From the correlation plot we understood that-

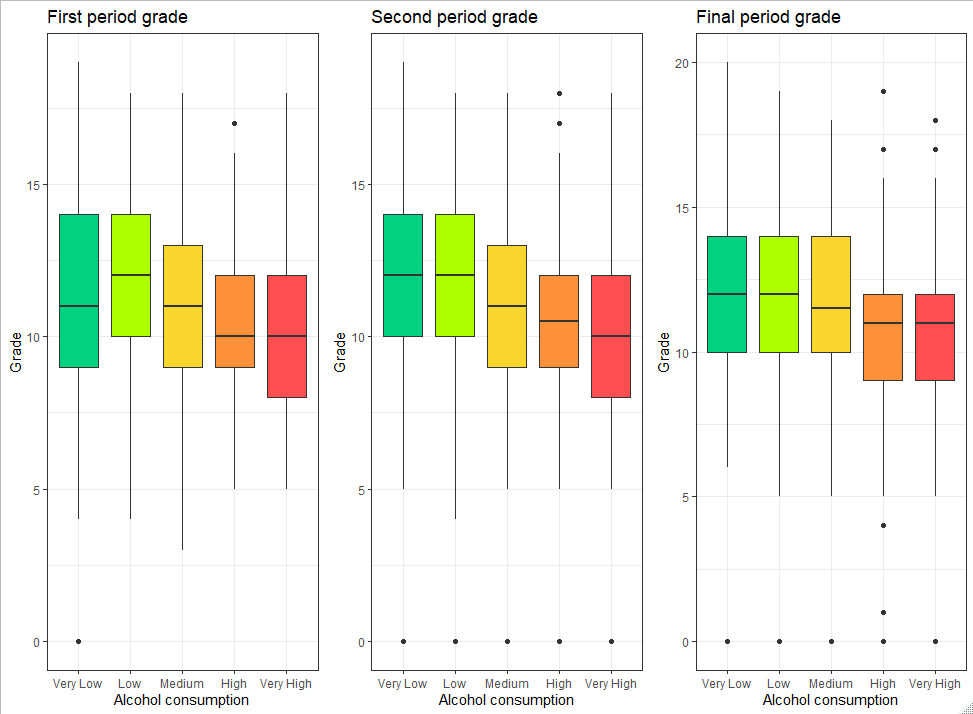
* There is a strong positive correlation among the grades. This could be because a student who performs well in first term is more likely to perform well in second term and final exam.
* There is a strong correlation between weekly alcohol consumption and daily alcohol consumption. This is because a person who drinks every week is more likely to drink every day.
* There is a strong negative correlation between grades and failures. This is because more failures will cause students to lose more grades.
* There is also negative correlation between weekly, daily alcohol consumption and study time. This is because a student who drinks every week is more likely to spend less time on studies.

**Data Visualizations:**



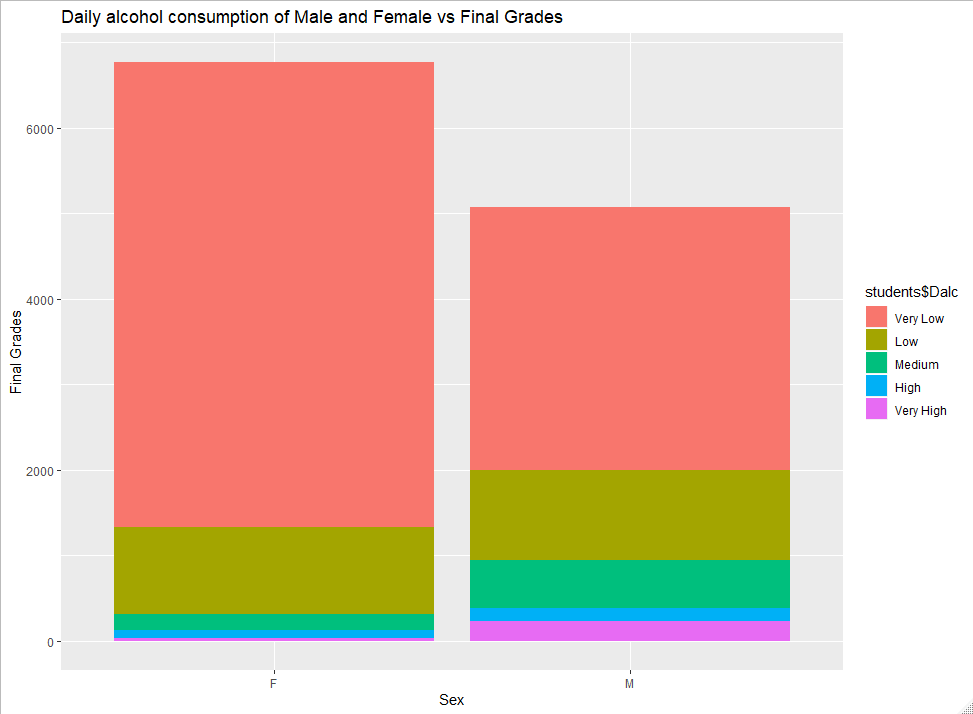
**Figure 3. Daily alcohol consumption vs grades**

The plot above is a comparison between the first term, second term and the final grade of the students who consume alcohol daily. We see that students who consume alcohol daily will most likely not perform well and have less grades. There are some students who consumed very little alcohol but got less grades and vice versa.



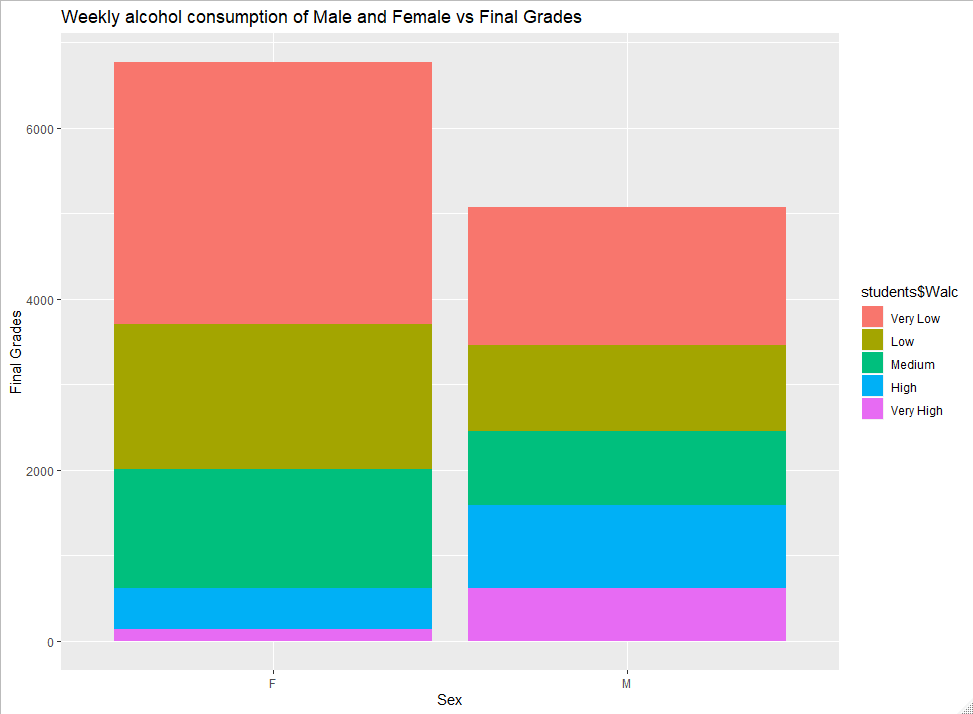
**Figure 4. Weekly alcohol consumption vs grades**

The plot above is a comparison between the first term, second term and the final grade of the students who consume alcohol weekly. We see that students who consume very low and low levels of alcohol weekly did better than the ones who consumed more alcohol. There are some students who consumed very high alcohol but got high grades and vice versa.



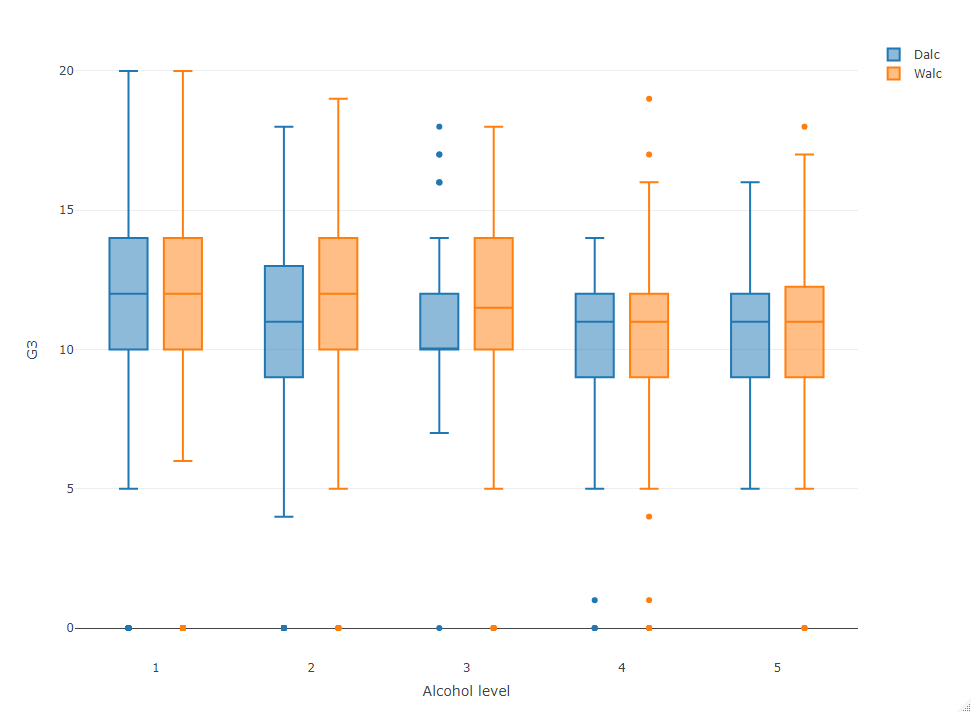
**Figure 5. Daily alcohol consumption of gender vs final grades**

The plot above is a comparison between the final grades and the gender of the students who consume alcohol daily. We see that students who consume very low levels of alcohol daily performed well in the final exam. Female students did better than male students and also consumed less alcohol.



**Figure 6. Weekly alcohol consumption of gender vs final grades**

The plot above is a comparison between the final grades and the gender of the students who consume alcohol weekly. We see that students who consume very low levels of alcohol daily performed well in the final exam. Female students consumed less alcohol and male students consumed very high levels of alcohol.



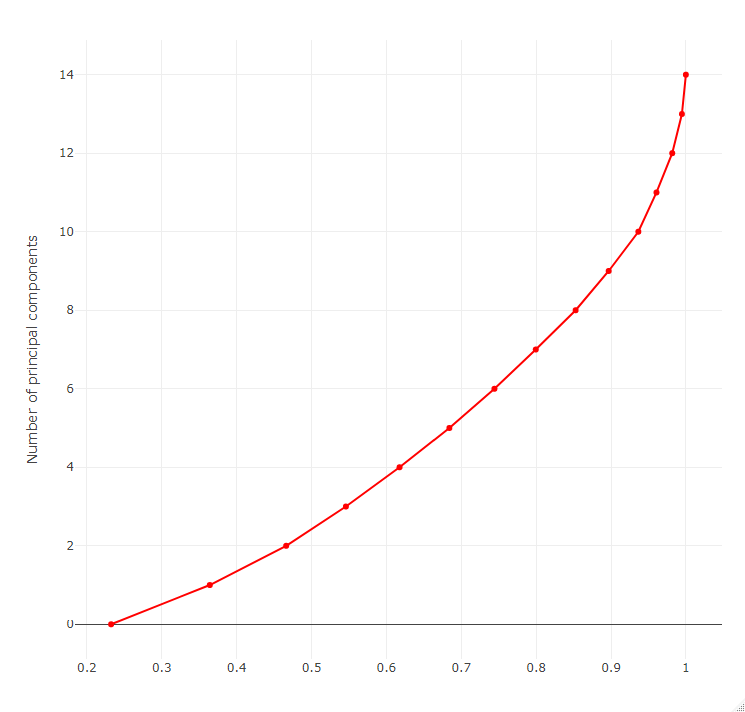
**Figure 7. Daily and weekly alcohol consumption vs final grades**

The above plot shows the daily and weekly alcohol consumption vs the final grades. We notice that people who consume alcohol daily have less grades in comparison to those who consume alcohol weekly. Some students are exceptional as they performed well even after consuming high levels of alcohol and vice versa.

**Principal Component Analysis:**

Principal component analysis will help us reduce the dimension of the data by considering only some important features in predicting the target variable. Despite having an overwhelming number of variables PCA helps us choose only the important variables. We can either do feature extraction or feature deletion to reduce the dimension of data. PCA helps us do feature extraction using which we can only consider significant features.

There is a significant increase in alcohol consumption during the weekdays and weekends. As analyzed, students prefer to consume alcohol mostly over the weekend, and the students with less grades have an alcohol rating equal to or close to zero. The main purpose of PCA is to find only a few features that can explain the variance in the dataset. PCA can cope up with real attributes and the attributes that have a natural order. We must exclude discrete attributes that don’t have a natural order, so we have removed 15 attributes. After analysis, we can see that the proportion of variance for 9 attributes is nearly 90%. Therefore, from PCA we can say that all the 15 variables are important, and we will continue to perform analysis with these 15 variables so that there is no information loss.

**Figure 8. Principal Component Analysis Plot**

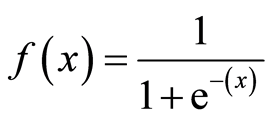
**DATA MODELS:**

For the prediction of final grades, we can use classification models. Classification can be done as the grades have been classified into 2 classes namely good grades and below average grades. Some of the classification models we plan to use are the Logistic Regression, Decision Tree, Gradient boosting and Random Forest. To implement these algorithms, we will do feature engineering to remove the highly correlated features and then proceed by fitting data into the algorithm. We will also analyze the confusion matrices to see how well the data models will perform and see which data model best suits our dataset. We will also understand how each algorithm works for our dataset. Using the varImp() function we can find out which feature is the most significant feature for predicting the grades. The accuracy report will also let us know the p value and other values which will help us better understand the importance of features.

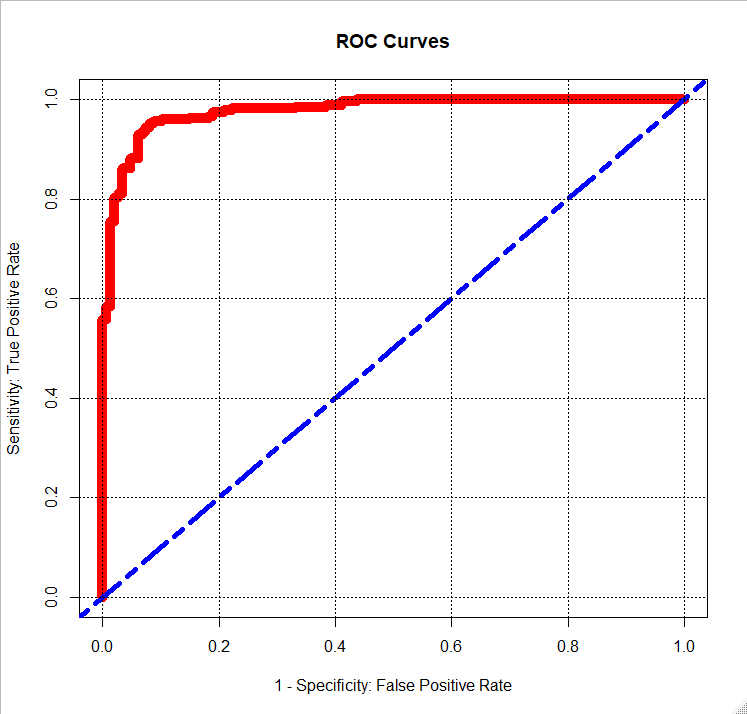
**CLASSIFICATION MODELS:**

**Logistic Regression:**

Logistic regression is a Machine Learning algorithm that is used in predictive analytics to classify data into different classes or categories and is based on the probability concept. There are two types in Logistic Regression namely binary and multi-linear class. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. The sigmoid function is used to map the value to its class. The formula is given as below.

****

We first implemented Logistic regression by considering all the 15 most important features and split the data into train and test sets**.** Sixty percent of the data was taken into the train and the rest into a test set. We fit the model and made predictions to get an AIC value of 287.56. To reduce the AIC value, we used the backward method in logistic as it gives us only the important features and improves AIC value. After using backward, the AIC reduced to 276 by only considering travel time, failures, first term grade and second term grade as important features. We got an accuracy of 94% for the new model. The ROC curve also shows that the model performed well with an AUC score of 0.976946. The closer that AUC value is to 1, the better our model performs.

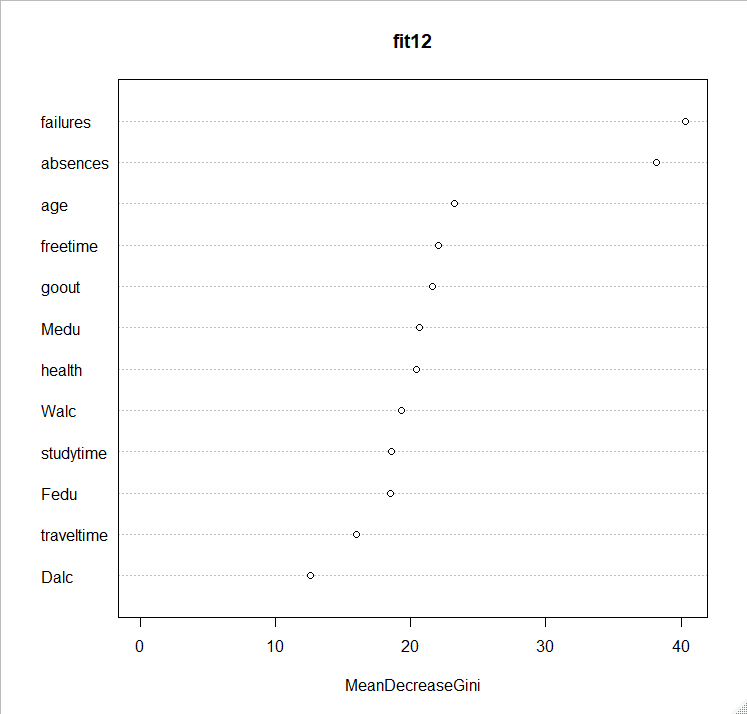
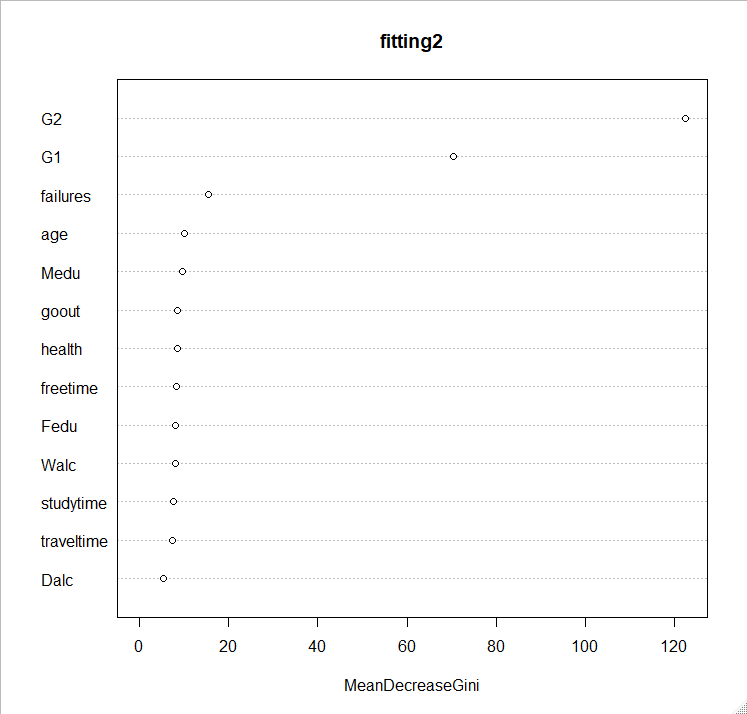


**Figure 9. ROC Curve**

**Random Forest:**

It is an ensemble tree-based learning algorithm. It generates a set of trees from a random set of training subsets. It aggregates the votes of different trees to classify the target variable into a class. These kinds of ensemble techniques use two or more algorithms to classify the target class. Random forest is known as a highly accurate classifier and it can run efficiently on large databases. It can also give us the significant features in a data set and can also estimate the generalization error. The mtry value will determine the number of splits on the data and changing its value can increase or decrease the performance of the model.

Firstly, we implemented Random forest on the 15 features that were significant. Then we used varImp() to see what the significant features in our data for final grade prediction were. Looking at the graph below, we see that first term grade and second term grade are the two most important variables in predicting the final grade. We got an accuracy of 93 % for this model with all features.



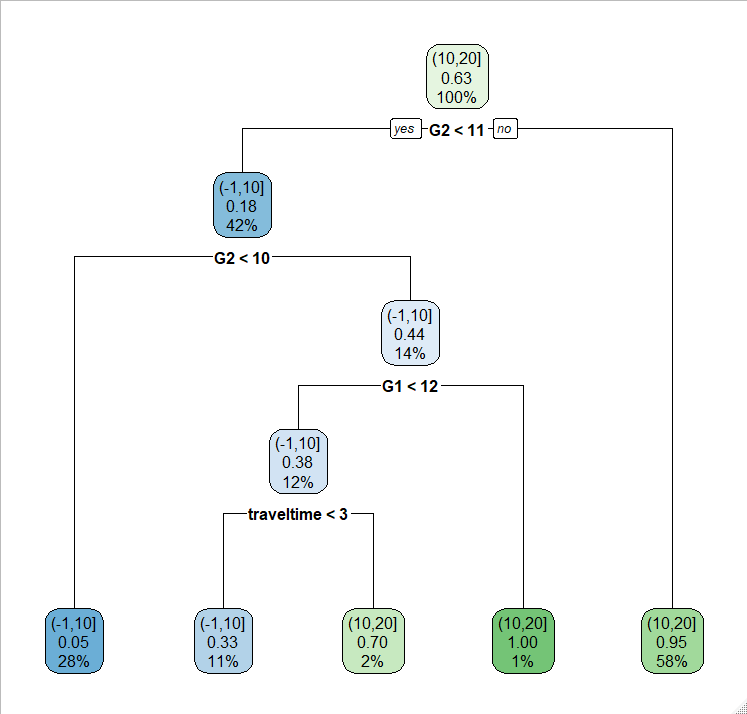
**Figure 10. varImp() for all features**  **Figure 11. varImp() excluding G1 and G2**

For the second part, we wanted to see how will the other features work in predicting final grade if we do not have the two most significant features, that is first term grade and second term grade, so we used the other features excluding these two and fit the data into the model again. This time the model did not perform that well. We got an accuracy of 71% and features such as failures,absences and age were proved to be the next significant features in determining the final grade. The ntree value was 500 and the mtry value was 3 in both these models.

**Decision Tree:**

Decision tree is a supervised machine learning algorithm which continuously splits the data according to the given parameters. The decision tree consists of 3 parts. They are the nodes, edges and lead nodes. The root nodes split the data according to a given condition and the output is given to the leaf nodes. The root nodes and leaf nodes are connected using the edges. Decision trees are split using the binary recursive partitioning technique. It is an iterative process of splitting the data into different parts and further splitting it until it classifies the data point. We choose the root node by calculating the Information Gain (IG). The feature with highest information gain is chosen as the root node.

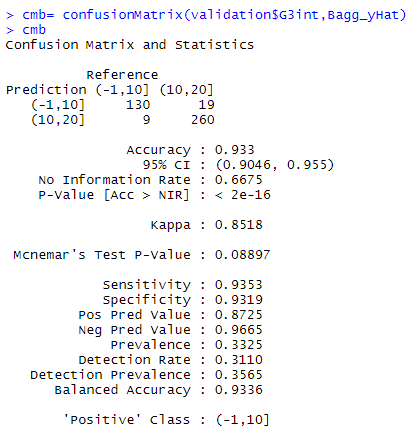
The decision tree below explains how our data was classified into good or bad grades. The root node here is the second term grade. It is the root node as it has the highest Information Gain. Students with grade less than 11 split into another node based on secord term grade again. This time it checks for grades less than 10 and splits the tree into subtree with root node as first term grade. The tree splits until no more classification can be done and the last root node is travel time. Here 61 % of the data is classified into good grades and the rest as below average grades. This can be noticed in the tree below.



**Figure 12. Decision Tree**

**Bagging:**

Bagging is another ensemble technique that is also called bootstrap aggregation and it uses multiple Machine Learning models to make accurate predictions better than a single model. It is used for data that has high variance. Algorithms like Decision trees do not work well for high variance data but bootstrap aggregation works well for such data. The algorithm works by creating random sub-samples of the dataset. It then trains each sample data and a new dataset is generated. The average accuracy of each model is considered to give out the actual accuracy of this model.

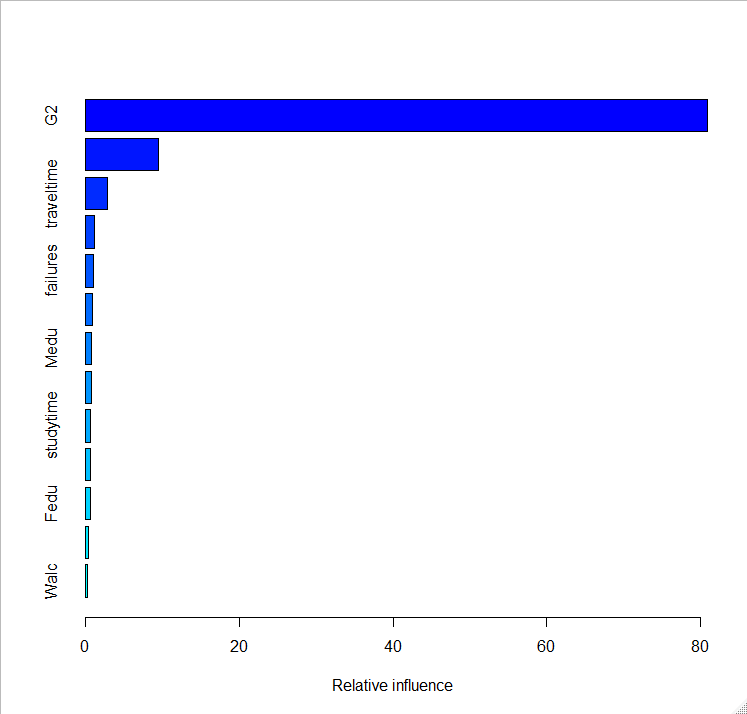


**Figure 13. Confusion Matrix and Statistics**

We performed bagging on our dataset, but this did not make much difference and achieved an accuracy of 93% with Kappa value 0.8531. We will perform our final model in the next step.

**Gradient Boosting:**

The idea of Gradient boosting algorithm is to convert the weak learners into strong learners. A weak learner is an attribute or a variable that performs slightly better than a random choice. To turn these weak learners into strong ones, we can use these boosting algorithms and Gradient Boosting is the revised framework of the original boosting algorithm AdaBoost. Gradient boosting works on three main elements. The first is to optimize the loss function. The next step is to use weak learners and make predictions. Using this weak learner to minimize the loss function is the last step. The four enhancements made to the boosting algorithm are tree constraints, shrinkage, random sampling and penalized learning.

When we implemented the gradient boosting algorithm, we got an accuracy of 93 %. The tuning parameter shrinkage was held constant at a value of 0.1 and the minimum number of observations in tree’s terminal nodes was 10. The final value of the number of trees was 100 and the number of splits it had to perform on the tree was 1. As we can see the plot, weak learners like mother’s education, studytime, father’s education and weekly alcohol consumption have now started to relatively influence the final grade. 

**Figure 14. Relative Influence of weak learners on Final Grade**

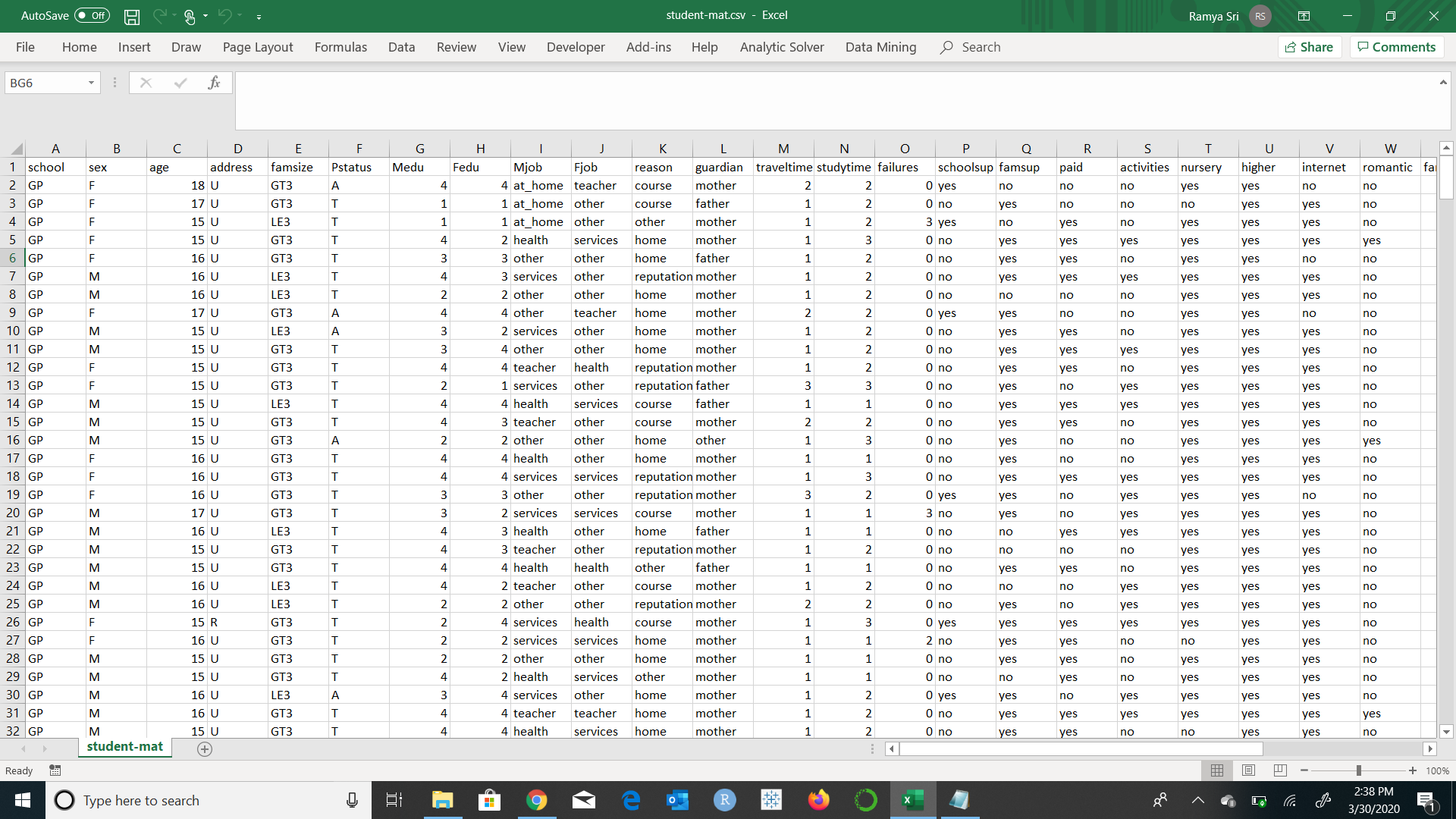
**EXPERIMENTAL RESULTS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification models** | **Accuracy** | **Error Rate** | **Sensitivity** | **Specificity** |
| **Logistic Regression(with all features)** | **93.54%** | **6.46%** | **0.9236** | **0.9416** |
| **Logistic Regression( with only 5 features)** | **93.77%** | **6.23%** | **0.9300** | **0.9418** |
| **Random Forest (with all features)** | **94.01%** | **5.99%** | **0.9127** | **0.9553** |
| **Random Forest (excluding G1, G2 features)** | **69.37%** | **30.63%** | **0.4899** | **0.8066** |
| **Decision Tree** | **93.54%** | **6.46%** | **0.9178** | **0.9449** |
| **Bagging** | **93.77%** | **6.23%** | **0.9353** | **0.9319** |
| **Gradient Boosting** | **93.77%** | **6.23%** | **0.9361** | **0.9386** |

**CONCLUSION:**

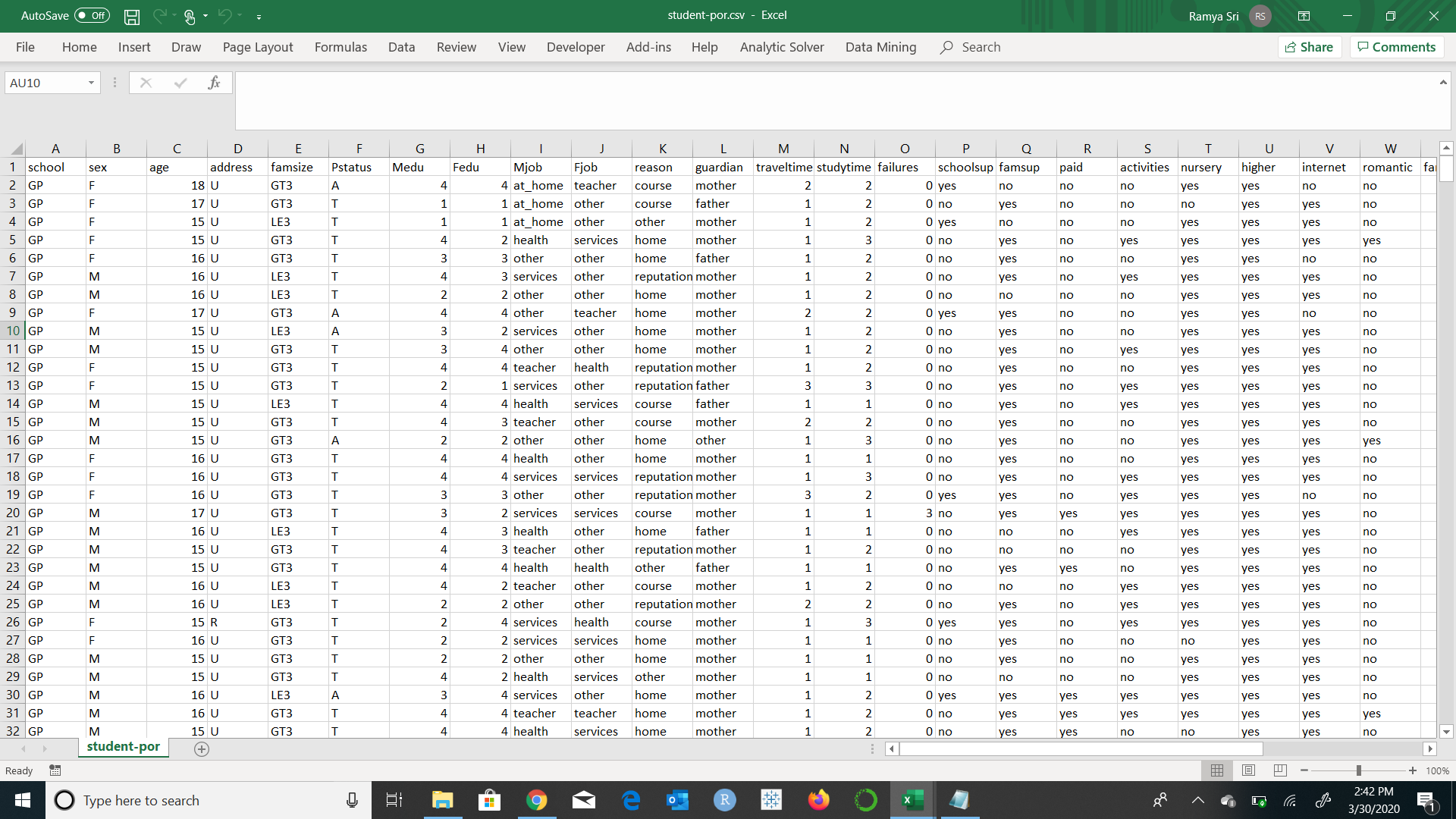
From the analysis, Alcohol is not only the one factor that affects the final grades of students. There are other features such as G1, G2, Failures, Absences etc which influence the final grades. From our analysis, we can see that, if students get good grades in first period (G1) and second period (G2), they will also perform well in Final grades (G3, target variable). But there are few students who have consumed high alcohol and have good grades in the final. All the models performed well except the random forest excluding features like first period G1 and second period G2 and among all the models Random Forest (considering all the features) performed better with accuracy of 94%.

**MATH DATASET SAMPLE:**



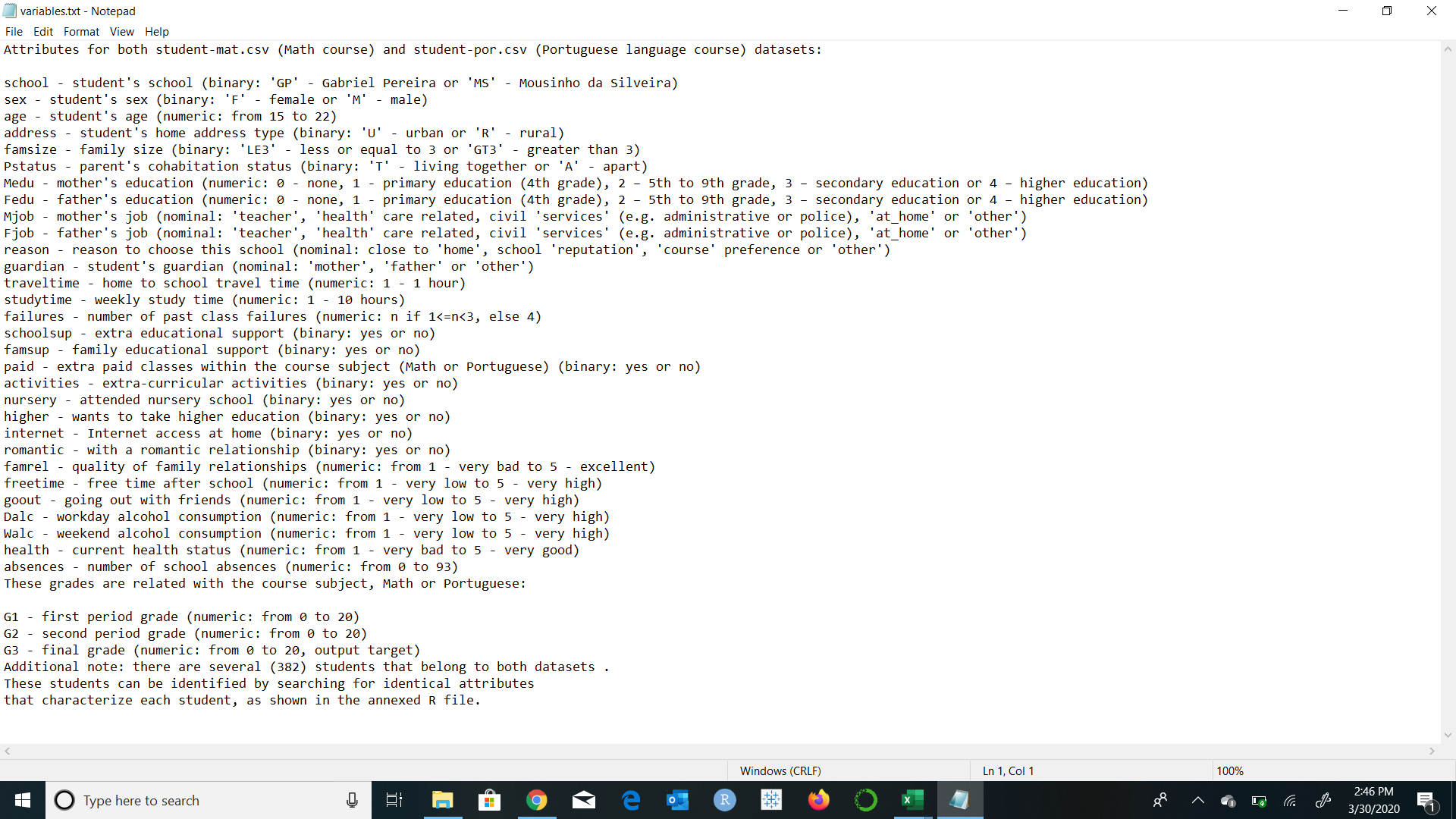
**Figure 15. Math Dataset**

**PORTUGUESE DATASET SAMPLE:**



**Figure 16. Portuguese Dataset**

**VARIABLE DESCRIPTION:**



**Figure 17. Variable Description**

**REFERENCES:**

1. *Student alcohol consumption*. (n.d.). Retrieved March 30, 2020, from <https://kaggle.com/uciml/student-alcohol-consumption>
2. chadha, c. (2020, January 2). *Bagging, Boosting and gradient boosting*. From towaards data science: https://towardsdatascience.com/bagging-boosting-and-gradient-boosting-1a8f135a5f4e
3. Koehrsen, W. (2018, August 30). *An implementation and explanation of the random forest in python*. From towards data science: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
4. Nagpal, A. (2017, November 21). *Principal Component Analysis - Intro*. From Anuja Nagpal: https://towardsdatascience.com/principal-component-analysis-intro-61f236064b38
5. Pant, A. (2019, January 22). *Introduction to Logistic Regression*. From towards data science: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
6. Yse, D. L. (2019, April 17). *The complete guide to decision trees*. From towards data science: https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14