# Title: Predicting Pass/Fail for High School Students

## Name: Neil Bhutada [Data Source: https://www.kaggle.com/uciml/student-alcohol-consumption]

Students from two high schools in Brazil ('Gabriel Pereira' and 'Mousinho da Silveira') participated in a survey where they gave personal and academic details. In this project, we will predict whether a student will fail or pass in school and understand various factors that affect the latter. These predictions will allow the schools mentioned (or other schools) to understand reasons for a student's failure and prevent the same by taking early precautionary measures.

For this project I combined two datasets. The first dataset had math scores and the other had Portuguese scores. The datasets are from Kaggle and have a "CC0: Public Domain" license. Both datasets have personal and academic information of the students, such as age, study time per week, father's education, mother's education, mother's job, father's job, rate of going out, free time per week, daily alcohol consumption, etc. There are 33 columns in both data sets. The first data set has 395 rows and the second has 649 rows. I combined both of these datasets by inner joining them on 19 columns (some of them being free time, study time, etc.). Then, I made a new column called "fail" that has binary outputs: 'True' and 'False.' I created this column by calculating the total math score and Portuguese scores; if either of the scores were less than 60% (benchmark for passing in Brazilian high schools), the value would be 'True' otherwise 'False.' The values in the "fail" columns will be predicted in this project. This final dataset has 20 columns and 162 rows.

By intuition, I thought that the 'rate of going out', 'amount of free time', and 'daily alcohol consumption' would be the most influential factors when it came to predicting if a student would fail or not. I thought that students who would fail will have higher values in all the columns mentioned above. All of these columns have ratings out of 5 (where 1 meant very low and 5 very high). Therefore, in figure 1, I plotted the average ratings of all of these columns based on a student's pass or fail status. My assumption about the 'rate of going out' and 'daily alcohol consumption' columns were right. The average rating for the 'rate of going out' for students who failed was higher by approximately 0.5 points. The average rating for the 'daily alcohol consumption' for students who failed was higher by 0.4 points. However, the 'amount of free time' had lower average ratings for students who failed by approximately 0.3 points. This can lead to inferences that students who fail have other obligations, such as part time jobs, to support themselves and their families.

But before training our machine learning model, I wanted to explore the dimensionality of my data across the 10 numeric columns by performing PCA. All the numeric columns (except for age, study time per week, time travel per week to school, and number of classes failed previously) had whole numbers from 1 to 5. We performed PCA on the numeric columns with and without scaling. We can see from figure 2 that without scaling we can capture about 95% variance within the first component! However, after performing standard scaling, we can see that we can't capture significant variance unless we use all the 10 numeric columns.

Finally, I trained my machine learning model to predict whether a student will fail. I used 19 features to train my model. The best model had a pipeline including: OneHotEncoder, PolynomialFeatures with n = 2, and LogisticRegression. This model had an accuracy of 76%, F1 score of 0.78, and AUC of 0.76. To make sure the model was not overfitting I checked the cross-validation scores and saw that variance was 0.007. Thus, due to the low variance, it can be inferred the model did not overfit. I tried other models as well (such as a model not using PolynomialFeatures, a model using StandardScaler, etc.). I also tried to stratify the training data against gender to make sure the training data set wasn't skewed for a certain gender. However, the models I got did terrible with AUC scores nearing 0.5. Thus, I refrained from using gender in my models.

From figure 3, it can be seen that the top 10 most influential features have come from the polynomial transformation. The squared of the rate of going out has the highest positive weight with around 0.267. This implies that having a higher score for 'rate of going out' will quadratically increases the chances of failing. The product of the free time with study time has the highest negative weight. This is because in the training dataset for most of the observations the study time is inversely proportional to the free time; the range is larger (from 1 to 5, while study time is 1 to 4) and variance is higher (by 0.2 points). Other crucial factors are age, number, mother's education, father's education, number of absentees, and daily alcohol consumption. Thus, the schools can focus on precautionary measures such as providing extra tutoring services for students who have do not have highly educated parents, alcohol consumption and awareness programs, planned study groups that would help students balance study and leisure time.

Figure 1: Average rating of Most intuitively influential column names with ratings out of 5 per pass/fail status

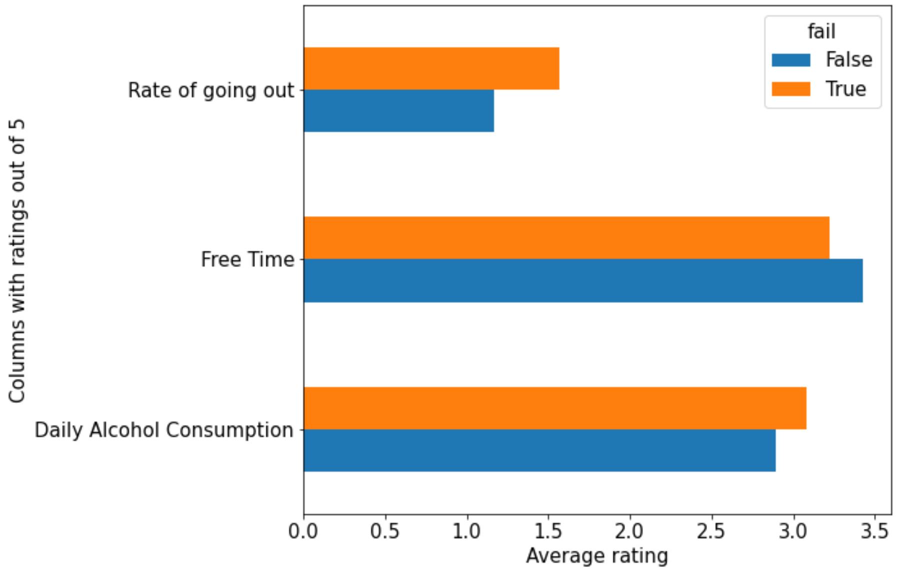


Figure 2: Cumulative Principal Components of Numeric variables

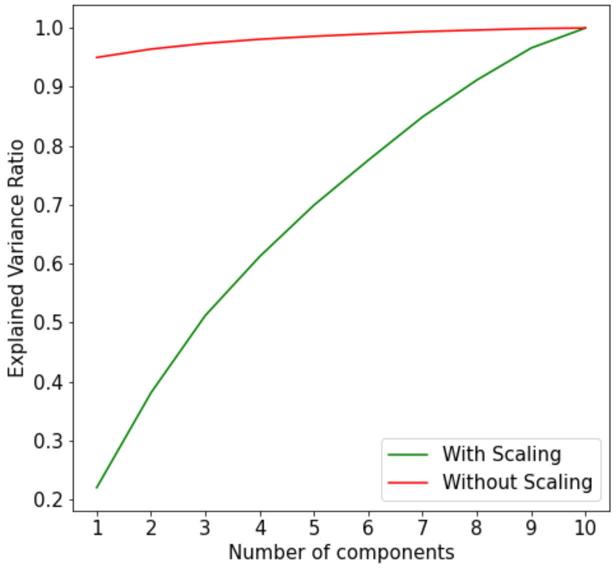
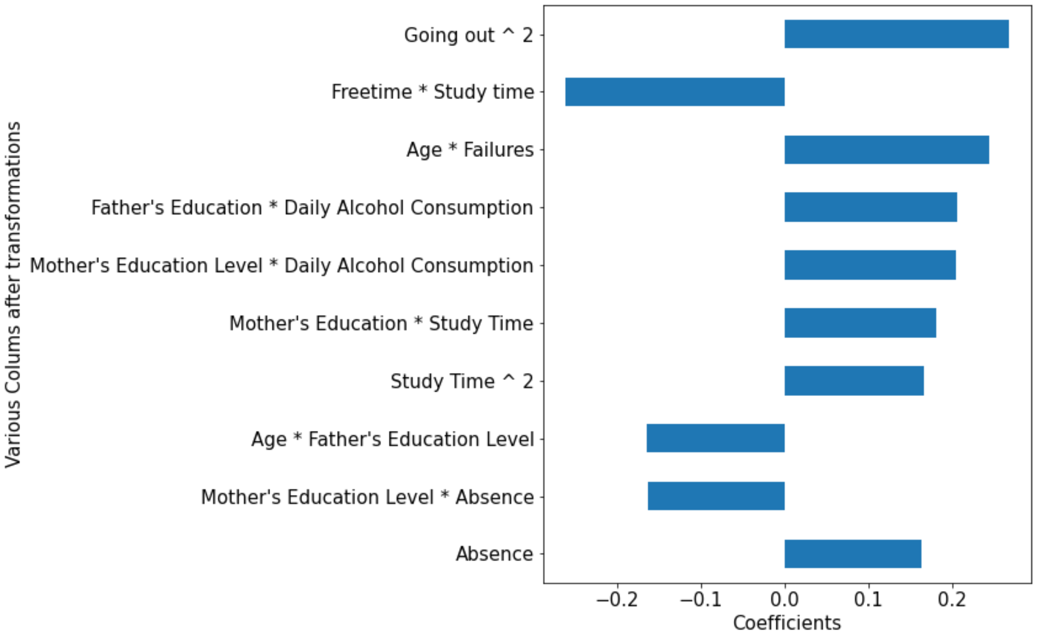


Figure 3: Top 10 most influential factors


Going out ^ 2
Freetime * Study time
Age * Failures
Father's Education * Daily Alcohol Consumption
Mother's Education Level * Daily Alcohol Consumption
Mother's Education * Study Time
Study Time ^ 2
Age * Father's Education Level
Mother's Education Level * Absence
Absence
Various Colums after transformations
Coefficients