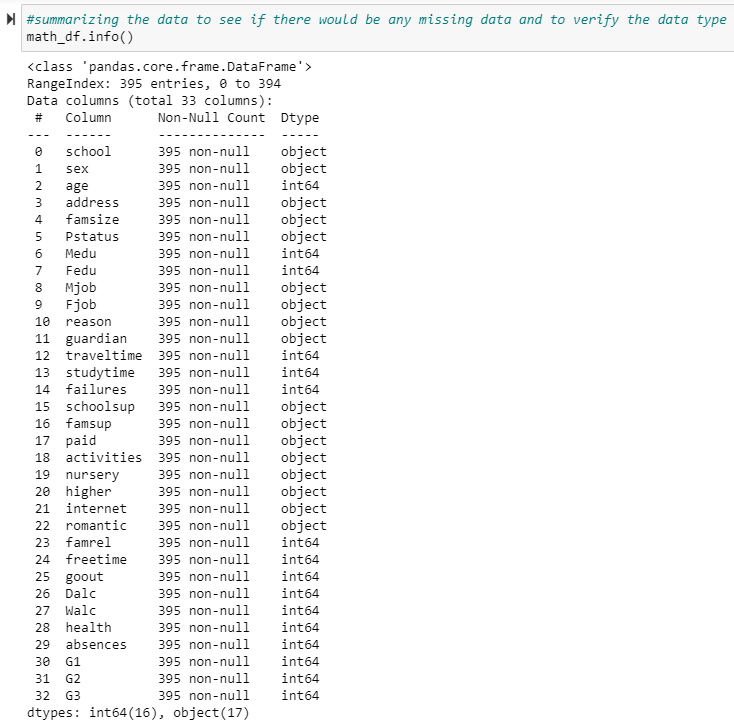
CAPSTONE 1 REPORT

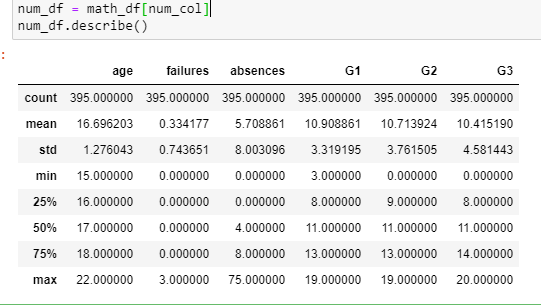
1. Introduction

In this CAPSTONE 1 Project, I am trying to predict Academic Performance in Mathematic for students from two schools. The potential clients for this research would be the schools or tutoring firms who would like foreseeing student’s performance. This could help the schools to take early actions to help student with potentially poor performance and/or provide a pool of good students to attend regional/national competitions. For tutoring firms, this could provide some good insights to identify business opportunities. The dataset I am going to use in this exercise is from UCI Machine Learning Repository ([https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance#](https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance)).

1. Data Wrangling

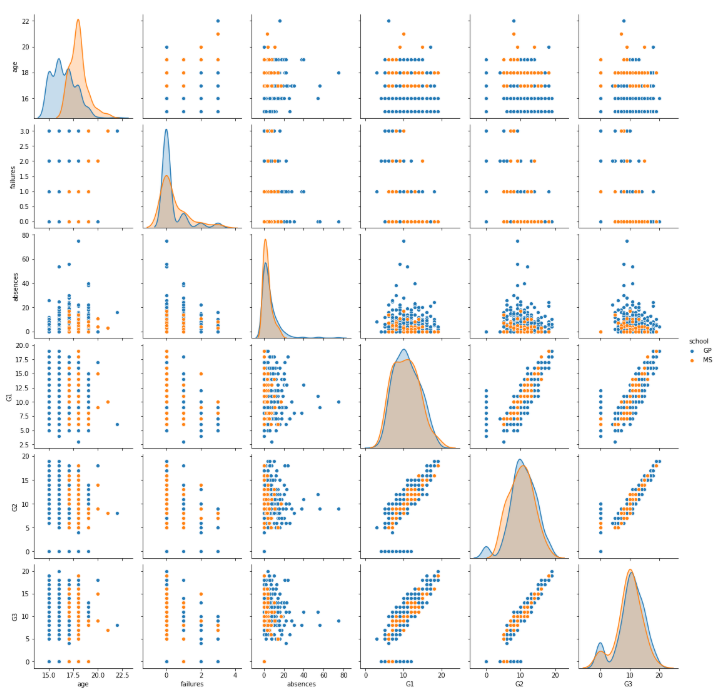
The data is in .csv form. It is loaded into python as a Pandas data frame. There are 33 columns and 395 student records. The target variable we are trying to model here is G3 which is the final score of students’ performances on mathematics. The data has been pre-processed from its source. Initial inspection shows no missing data.

Reviewing the dataset indicates that even though there are columns with numerical values, but they are actually categorical. Therefore, we will make separate groups of columns (numerical-num and label-lab). Numerical columns include ['age','failures','absences','G1','G2','G3']. Categorical columns are ['school', 'sex', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu', 'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health'].

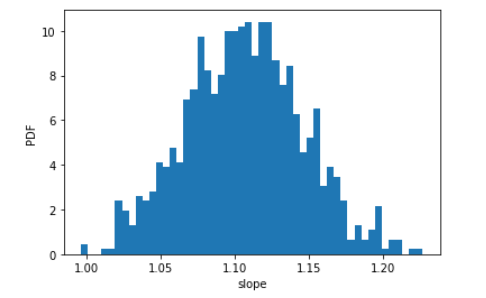
We will also explore some statistics of the numerical variables

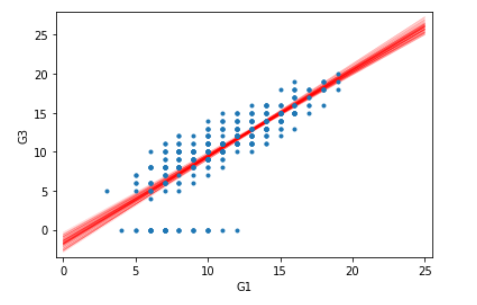
1. Exploratory Data Analysis

Box plots does not indicate significant or unreasonable outliners. Pair plots indicate some correlations we should look at.



It seems that absences and performances on the first and second exams (G1, G2) have influences to the final grade (G3). Furthermore, the first and second exams are also correlated. Students who performs well in G1 tend to perform well in G2 and G3. The correlation matrix indicates that the other predictor variables are independent on each other. There are different distributions for some variables (school) but the violin plot does not signify or much different in terms of correlation to G3. We performed pair bootstrap to see how the correlation between G3 and G1 and G2 would change. It was concluded that the correlation is stable given the data set.





We can clearly see that linear regression cannot model of the data which is not outliers. In addition, as we need to involve other categorical variables, we can consider other tree base prediction algorithms.

Statistical analysis by eyeballing, we see that the two school has almost identical G3 and Absences distribution. Statistical analysis is also performed to test this hypothesis. There are 19% we would get difference in mean of G3 for two schools or 81% we can the same mean of G3 for if we model two schools together.

1. Prediction

Data is split to train, valid and test the model. 70% of the data is used to for training and validation, 30% will be used to test the model. Random Forest and Gradient Boosting are used in this prediction. Both do not show significant difference in predictive power during training, parameter tuning and validation. However, testing shows that random forest performs a bit better in this situation (88% versus 85%). Random forest also shows that G2 and Absence are the two significant variables driving G3.

