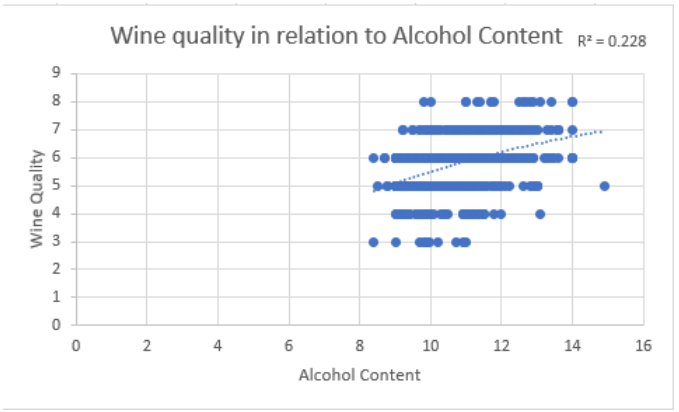
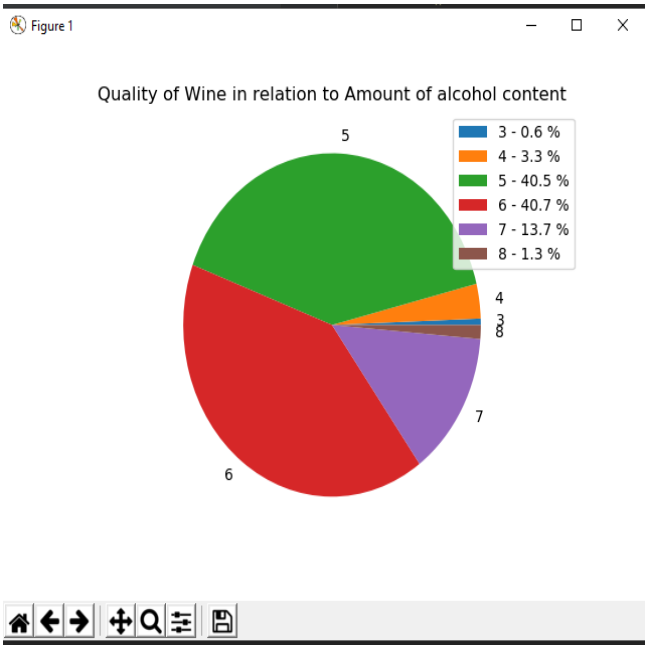
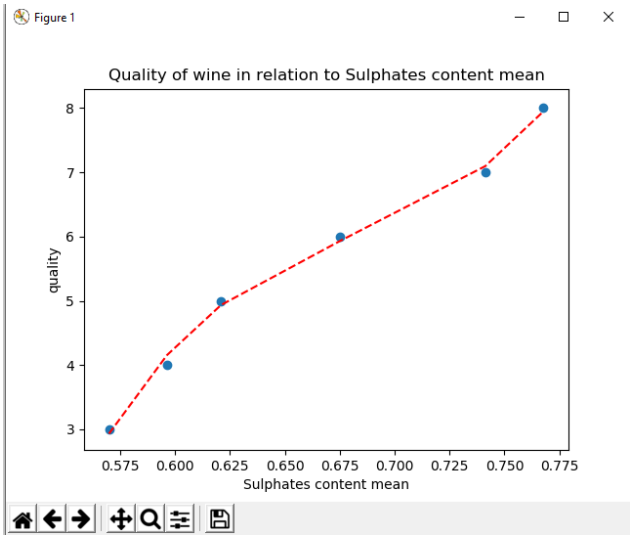
Report for Machine Learning Final project

My project the dataset I have selected Is *Wine Quality Data Set* created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis. This Dataset contains 1,600 entries of information separated over 12 attributes. Most of the attributes refer to the content of the ingredient within the red wine and the last attribute depicts the “quality” of the red wine. The red wine is a specific brand by the name of "Vinho Verde" which is Portuguese wine. The angle I decided to approach this dataset by is trying to use Machine Learning Algorithms to predict the quality of the wine based on the correlations/trends it learns from studying the attributes in the training set in relation to their class of quality. This dataset could be looked at from numerous perspectives such as trying to predict other attributes in the dataset, however the Quality class is the best since it has recurring values which makes it easier to classify the wines. Predicting is one thing, but how does this data relate to the real world, well in the real-world companies wishing to produce better quality of red wine could use this data alongside some machine learning algorithms to best correlate the right amount of ingredients required for the best quality of red wine. By predicting what ingredients affect quality in what manner the company will then be able to find the sweet spot where quality and price for ingredients meet so that they are able to maximize their profits and reduce losses when releasing a new Variation of the red wine. Furthermore, companies could also use the data and figure out what attribute has the smallest impact on the quality of the wine and decide weather its worth the cost to even incorporate in the wine. Overall, the data set is pretty much a normal distribution where majority of the wines are of average quality and bad/good quality wines are outliers.







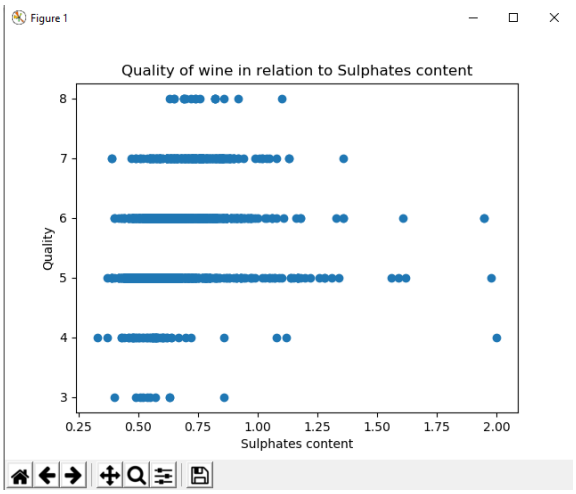
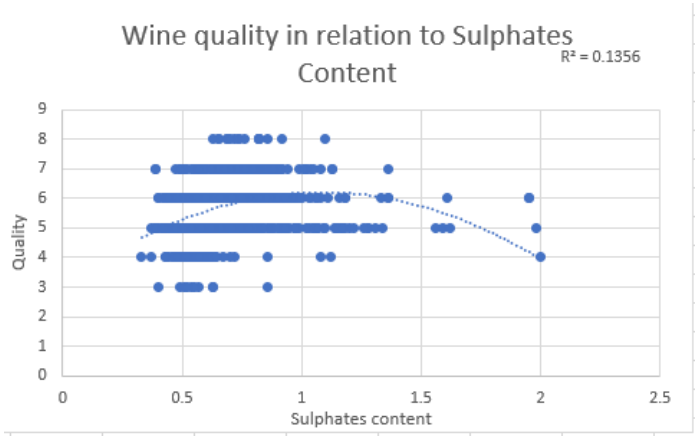


Figure 1
Quality of wine in relation to Sulphates content
Quality
Sulphates content


Wine quality in relation to Sulphates Content
R² = 0.1356
Quality
Sulphates content

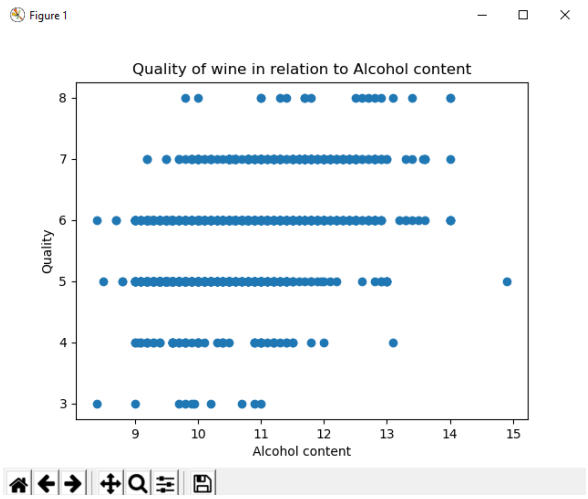


Figure 1
Quality of wine in relation to Alcohol content
Quality
Alcohol content

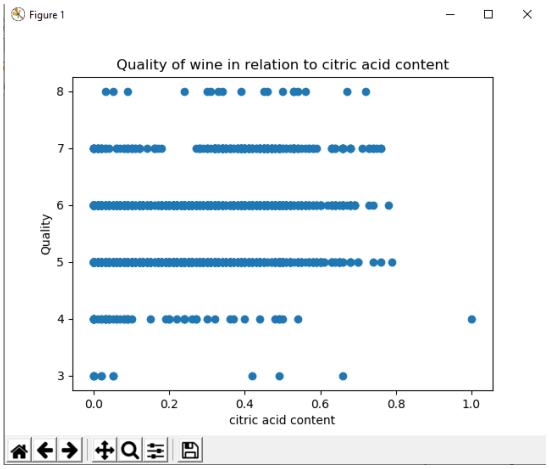


Figure 1
Quality of wine in relation to citric acid content
Quality
citric acid content

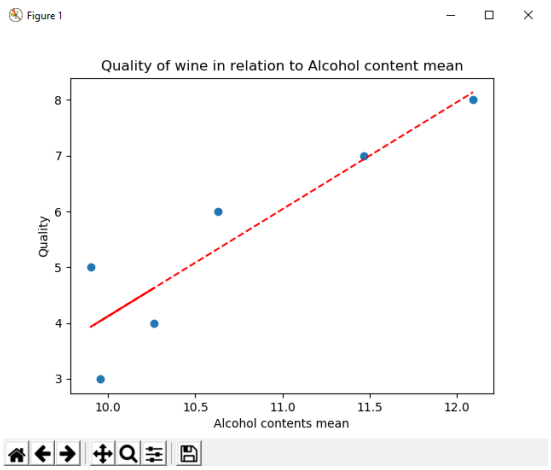


Figure 1
Quality of wine in relation to Alcohol content mean
Quality
Alcohol contents mean

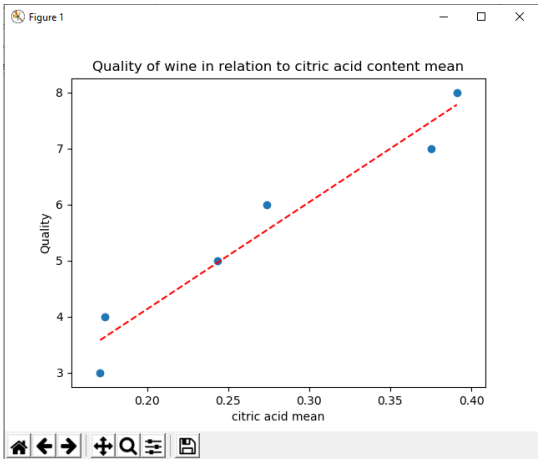
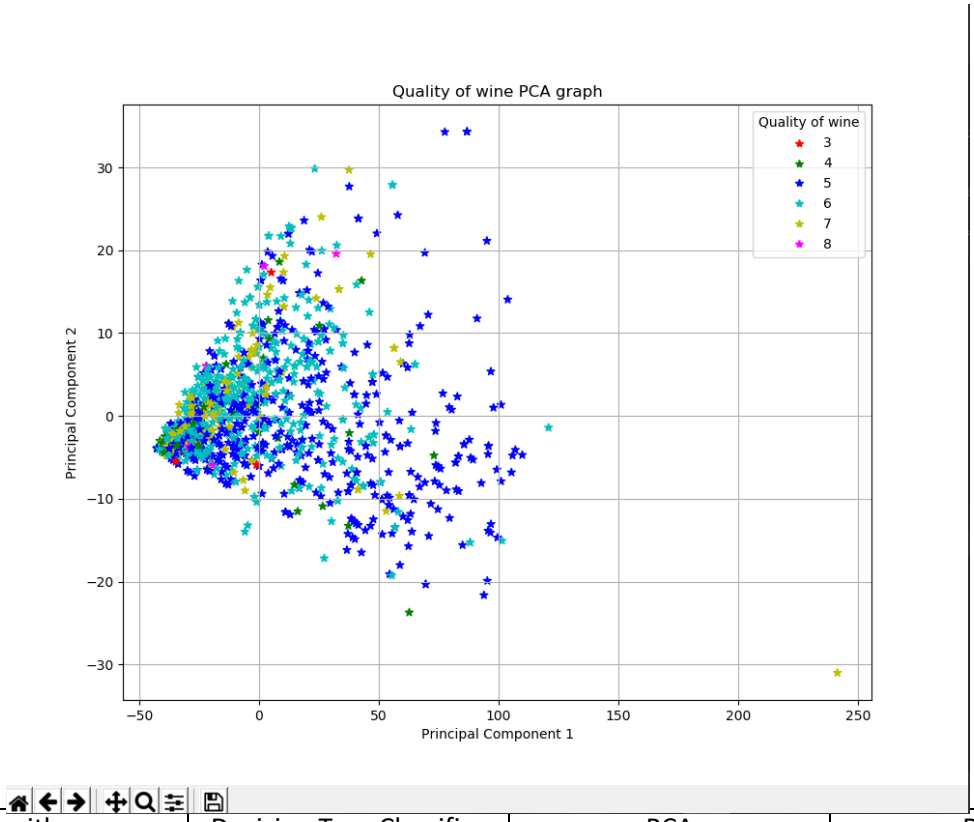


Figure 1
Quality of wine in relation to citric acid content mean
Quality
citric acid mean

Preparation of the data:

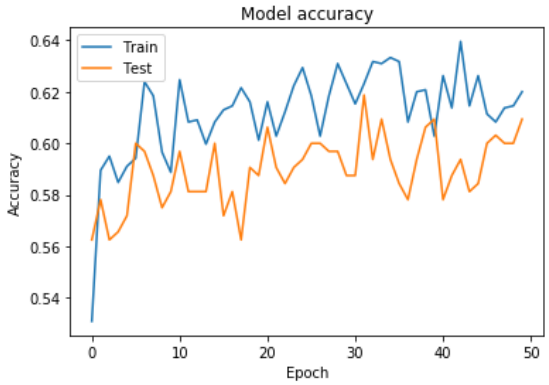
For the most part, the data was pretty much usable, however after doing a decision tree classifier and printing out the correlation matrix for each attribute it was clear to see how some attributes affected the quality very minimally thus skewing the prediction analysis. To prepare my data for the algorithms to better predict the quality of the wine, I dropped an attribute when training and testing my data. The attribute I dropped was “citric acid”, my reason for dropping this attribute was because it proved to have the smallest correlation to the wine quality and thus was skewing the training and predicting process. In addition to dropping an attribute I also had to convert the quality column of the data frame from type int64 to type String/Object, the reason for doing this was so that my Decision Tree Classifier would accept it as a list of classes and then create the binary tree with the proper classifications. Lastly, I tried my DNN from multiple perspectives one of which being dropping all the wine entries with qualities not equaling 5 or 6, by dropping every other quality of wine the DNN was then able to better predict the quality of wine for a multitude of reasons. The DNN was better able to predict after the modification to the data because not only was it now prediction binarily but it was also supplied with ample data for the only classes remaining thus making the training and predicting accuracy much higher.

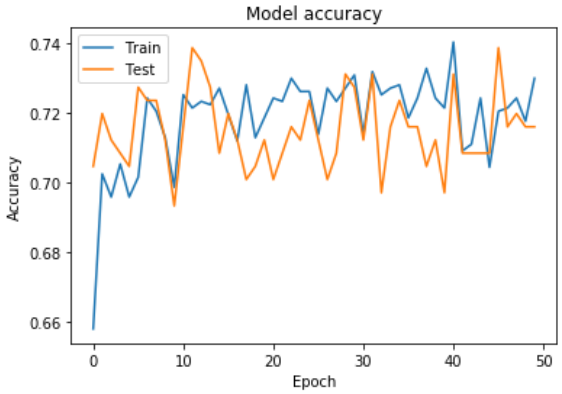


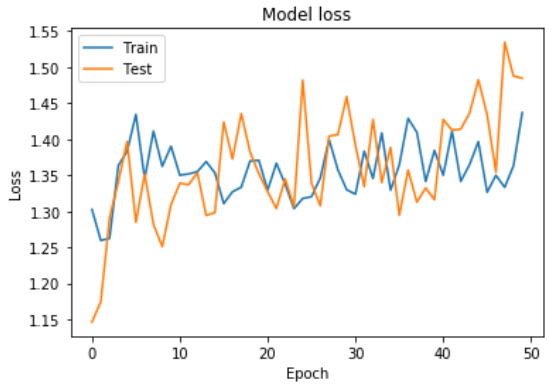
Algorithm:	Decision Tree Classifier	PCA	DNN
Accuracy:	65%	N/A	78%

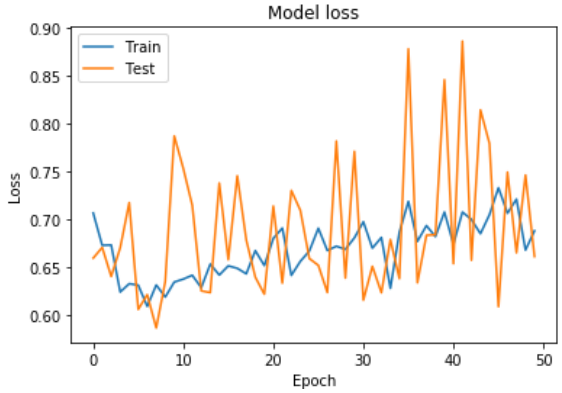
The 3 Algorithms used for this project were, Decision Tree Classifier, Principal Component Analysis, and Deep Neural Network. When Comparing these 3 algorithms performance it was a bit tricky since PCA did not perform any kind of prediction on its own and therefore was difficult to predict the performance of it. When doing the Decision Tree Classifier, I had to take in to account what attribute would be the best to classify and from that create a decision tree. The decision tree plotted the classifying excellently for the training data however, when measuring the accuracy of the prediction using the Decision Tree Classifier, much to my surprise the prediction accuracy was not nearly as high as I expected it to be, it averaged around low to mid 50s %. When training and testing the accuracy of the Deep Neural Network I ran many different tests to see how it would affect the accuracy of the algorithm. When conducting these various tests on the Deep neural network I realized that when predicting different columns my accuracy would vary heavily than when predicting a different column. During these tests I trained DNN with numerous different labels such as citric acid (which showed as having the weakest correlation when creating the decision tree), alcohol content, quality and sulphates content. When training to predict citric acid, I noted that the performance of the DNN was horrible yielding a low accuracy of 7.8% and a high accuracy of about 8.5. Clearly due to the weak correlation of citric acid to the rest of the data the DNN had a hard time learning to predict a correct citric acid based on the rest of the data. When training to predict sulphates content and alcohol content the DNN displayed a noticeably large difference in accuracy than when predicting citric acid, and rightfully so. Lastly, I predicted quality of wine based on the data without the citric acid column since it had weak correlation. When predicting the quality based on the data my DNN produced notably the best result between my other DNN simulations but also between the three algorithms yielding a highest validation accuracy of about 70%. I figured DNN predicting class was the best option however, even when predicting quality, I seemed to not be able to get above 70% validation accuracy how I could make my DNN even better. Finally, I realized that the data was in the form of a normal distribution where most of the data was on average quality wines (quality 5, quality 6). After realizing the fact that the data was in the form of a normal distribution, I decided to drop all other quality entries and predict based on just quality 5 and quality 6 wines and as my hypothesis suggested the validation accuracy did in fact rise, all the way to a high of about 78%. For my PCA algorithm, I tried to fit my data that was reduced through the use of the dimensionality reduction algorithm PCA into a SVM model and also into a linear Regression model however, much to my disappointment failed in doing so, thus resulting in my PCA not being able to predict the data. After fitting the PCA data to the SVM model and trying out predictions my code would get stuck at the predicting phase and not be able to continue further to print confusion matrixes and classification reports. Overall, it was a close call between Decision tree classifier and DNN however, DNN took the cake when predicting the quality of wine. The Decision Tree Classifier did however, get very close resulting in an all-time high accuracy of about 65%.

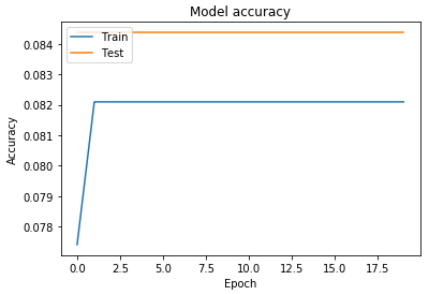
Deep Neural Network Graph when I removed all qualities aside from 5,6 since they had the most data on them. As you can see for these graphs the accuracy shoots up since its guessing binary values 5/6 == 0/1 of which it has ample data for. The figure shows the test accuracy hovering around 70.5%-74%~(Right)











These Graphs show the Model Accuracy/Loss when predicting all 6 different wine qualities. As the graph depicts the accuracy when having to predict 6 different classes when lacking ample data for each of the classes is significantly lower than when just predicting for wine quality of class 5,6.(Top Left)

This Last Graph to the left depicts the accuracy when attempting to predict the citric acid content in the Red wine. As the graph shows the accuracy caps out around 8.5% this evidently proves the weak correlation Citric acid has with the rest of the data set and should be dropped when predicting the quality of the Red Wine.(Bottom Left)