Predicting Wine Type and Quality

Pratik Patel

Introduction:

Wine has been around in the world ever since 6000 B.C, with the oldest known winery being at least 6100 years old. Unlike some other alcoholic drinks which have a negative connotation such as beer, wine is one of few alcoholic beverages that continues to have a positive connotation mostly with it being an elegant beverage for the higher class. However, any normal wine is not considered elegant but wine that has aged for years which supposedly enhances the taste and quality of it. In the real world, individuals who can notice the slightest of difference in the quality of wine, known as wine connoisseurs, are deemed the most sophisticated. More and more people are drinking wine and trying to become oenophiles by going to events like wine tastings. While humans use their sensory feature of taste to determine quality of the wine I would like to see if the scientific features that make the taste better could be used to make a computer "connoisseur".

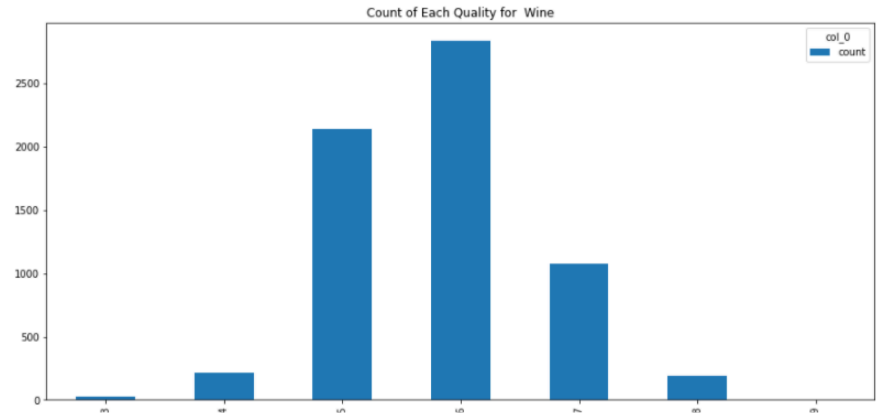
We are going to see if some features such as the pH level, citric acid, alcohol, and other features can determine what makes the quality of wine better or not. Since wine comes in different variants I will see if first by using all the 12 features I can predict the wine type as in the dataset there are two types of wine red and white. The more difficult task will be once again using the 12 features see if we can predict the quality of the wine. In a sense, our computer connoisseur will be able to predict the wine type and quality of the wine like how an oenophile can. Another objective of the lab, will also to see which supervised learnings algorithms does the best job of classifying the type and quality.

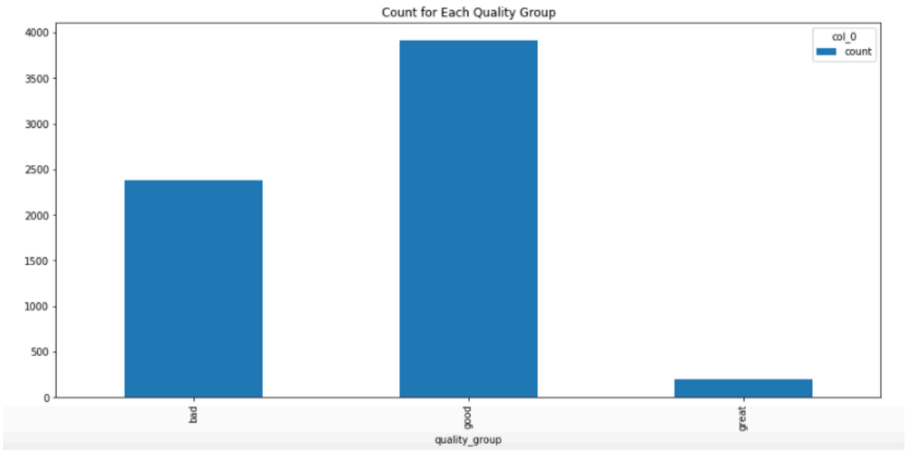
Methodology:

The data required some preprocessing which dealt with first looking at the separate data of each variant to see if the there was some correlation between the features. Since there was very little correlation I decided to join the two data tables into one large dataset. Before joining, I added a column in each type (wine and red) which was just a label of the type of wine. After performing an EDA to see the count of each quality, I noticed that the range of values was not great as it only went from 3-9 with majority of the data towards the middle which then later caused error when tried to predict. To fix this, I created another column in dataset which basically turned the 10 supposedly categories to 3 where if the quality was less than 6 it was deemed bad, between 6-7 was good, and anything greater than 7 was great to help train a better predictor for quality.

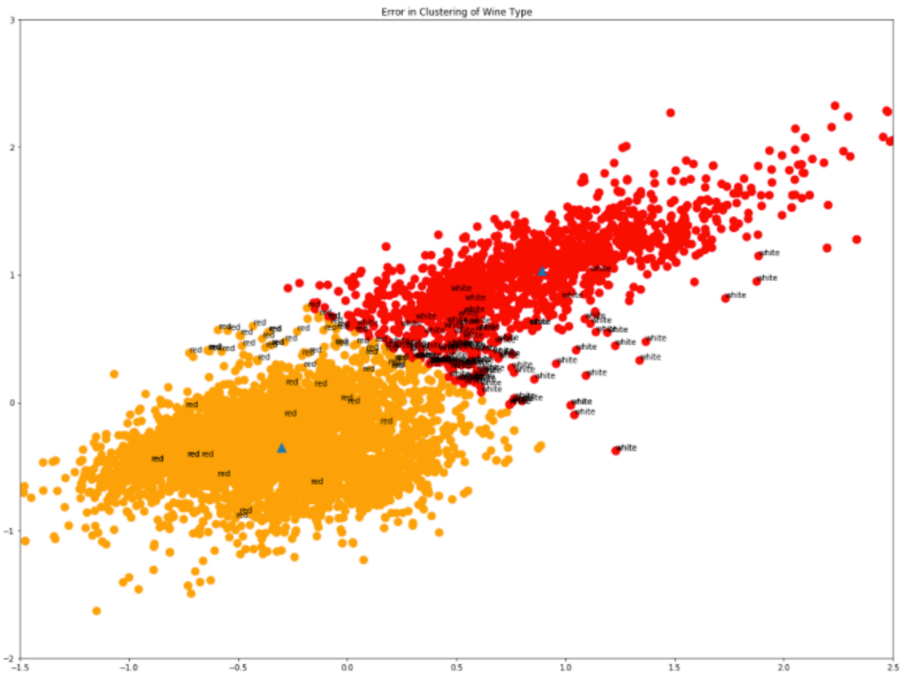
The dataset was shuffled, as originally it was sorted by type, to make sure randomness was even more ensured for when I have to split the data to testing and training data. Before actually trying to train a predictor, the data needed to be scaled as it is one of the major requirements of any supervised learning algorithm and therefore I used sklearn StandardScaler class to accomplish task. I also used the sklearn LabelEncoder to convert both the wine type and the quality groups to numerical values. With all the data ready now to be trained I used different supervised learning algorithms using mostly default parameters such as KNN, a decision tree, Logistic regression and etc. as well as created a deep neural network. Specifically, for wine types, I used SVD to reduce the dimensionality then cluster the data to see if clusters are formed for each type of wine. To check accuracy, I used the score method as well as looked at the confusion matrix to see how it was being classified.

Result


Count of Each Quality for Wine


Count for Each Quality Group

The EDA done showed us that many of the features do not really correlate to each other and when they do they are features that will relate but do not necessarily mean it easier to classify. Overall, from the bar graphs of the count of each quality measurement we can see how the data itself is not that good as most of the data is in the range of the quality being between 5 and 7 when then entire range is from 0-10. Even when I split the data into 3 categories oh bad, good, and great it is seen how the data is severely unbalanced with there being very little high-quality wines (8,9,10) which is the reason why accuracy is not great for predicting quality later discussed. For the wine types, we do have enough data for each types as it is only a binary classification problem with 25% data being red and the other white. When I used k-means to cluster I saw that red and white wine do indeed create good clusters with there being some overlapping part causing error in clustering seen from the plot where the data points with the text are the wrongly clustered point according to wine type. Overall, for the wine type classification we can see that it was quite easy to predict the type as mostly all the predictors gave an accuracy of 99%. The results for the classification of wine quality was not great as the neural network



gave only a 72.4% accuracy and out of the other classifiers used the highest was still only 74%. The error came from how the dataset was not balanced in the sense of not really being much data on high quality wine and therefore by looking at the confusion matrix we can see how the predictors rarely made any guess' classifying a wine with high quality. Overall, I am happy with the wine type classification however disappointed in the wine quality classification as I could have improved it by getting more data if there was more time and would also like to create a classifier with the data being on one type to see if different type affects quality differently. (For more graphs and tables go to .ipynb file)

Conclusion:

In conclusion, we know that different types of wine have different scientific properties which allowed us to make a highly accurate classifier for wine type, but not for wine quality as the quality. While, it was not the greatest classifier we do know that wine quality still is dependent on the 12 attributes in some way as we still get a 74% accuracy.

References:

1. http://docsdrive.com/pdfs/ansinet/jas/0000/42412-42412.pdf
2. http://rstudio-pubs-static.s3.amazonaws.com/24803_abbae17a5e154b259f6f9225da6dade0.html
3. http://rpubs.com/garrym3k/175762