Quality of Good Wine

Name: Shourya Agrawal

Wine, being a food, is very subjective in nature. While Person A might be fond of a certain wine, Person B might not think the same. For my project, I have tried to demystify this flavor profile of wine into more objective numbers. I tried to experiment on which factors change wine for the better, and which for the worse. This can potentially be used by distillers, or wine enthusiasts to estimate how good their product will be.

This dataset was published by Paulo Cortez, University of Minho, Portugal, to the University of California – Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/wine+quality. This data set consisted of 3 files, 2 of them being csv files of red and white wine respectively, along with 4899 x 12 for white wine, and 1600 x 12 for red wine.

First, I checked the data, if it had to be cleaned. There were no missing values so we moved on ahead. Now, lets get the basic idea about our data. Figure 1 is a stacked bar plot showing the number of wines belonging to each quality group, and how many of those are red / white. This was done by first allotting type to them (red / white) and then merging the 2 data sets. By far the most common quality of wines is 6 (43%), with white being the majority here. Red only takes up 25% of the 6 plot here and even lesser is cases like for 7 and 8. Even in general, white wines exceed reds ones, but this is mostly because the dataset for white was larger. This, along with another reason I will discuss later, is why being red / white will not be taken into account while deciding which wines are good. We can see '9' quality wines are far to less to even be represented.

Since all of our data is numerical, we can not start columns from consideration unless we know that they are not of use, and do not corelate to being good or bad. In my caser, I am defining good wines to be of a quality > 6. And bad to be < = 6. I am using all the columns (1) fixed acidity, (2) volatile acidity, (3) citric acid, (4) residual sugar, (5) chlorides, (6) free sulfur dioxide, (7) total sulfur dioxide, (8) density, (9) pH, (10) sulphates and (11)alcohol.

Now, we will experiment with the dimensionality of our data. We have 11 features as of now, not including the type that was added by me earlier. And we shall be preforming Principal Composition Analysis on these. Although the figure does not show it, I found out that not scaling the data had about 98% for explained variance just for a single component, which is not ideal. After scaling, we see that we can explain about 87% of variance for Red Wines, and about 82% for White wines.

Lastly, Figure 3 shows coefficients of the most important factors which corelated to the quality of our product. For that, I tried several things. Firstly I tried linear regression for all, followed by polynomial regression of all. Linear worked best, so I used that. Then using a map of correlation, I found the most useful factors for both red wine and white wine. Red wine had more clear factors, and just used 6. Its accuracy came out to be 89.5%. White on the other hand used 8 and still had only 80% accuracy. Since the factors affecting both were different, we could not have combined both into one and still received such good results. I did try that though ands got about 80%, the same as of white.

In conclusion, taste is not as subjective as initially believed, and can be somewhat efficiently calculated

Figure 1: Wine Distribution

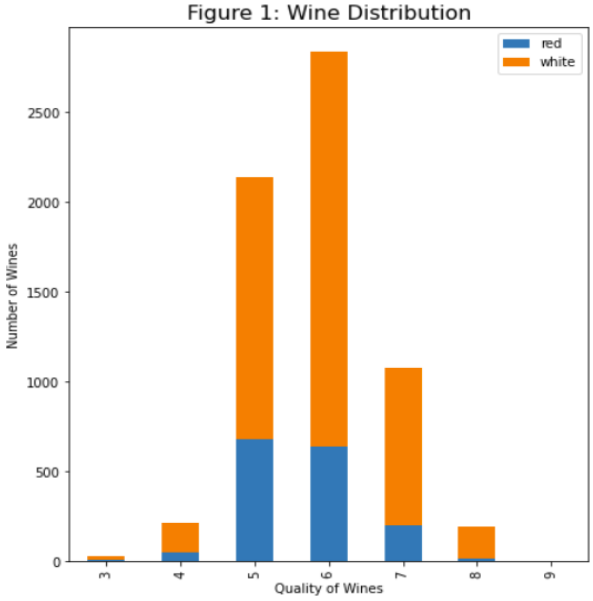

Figure 2: Cumulative Principal Component Analysis (Scaled)

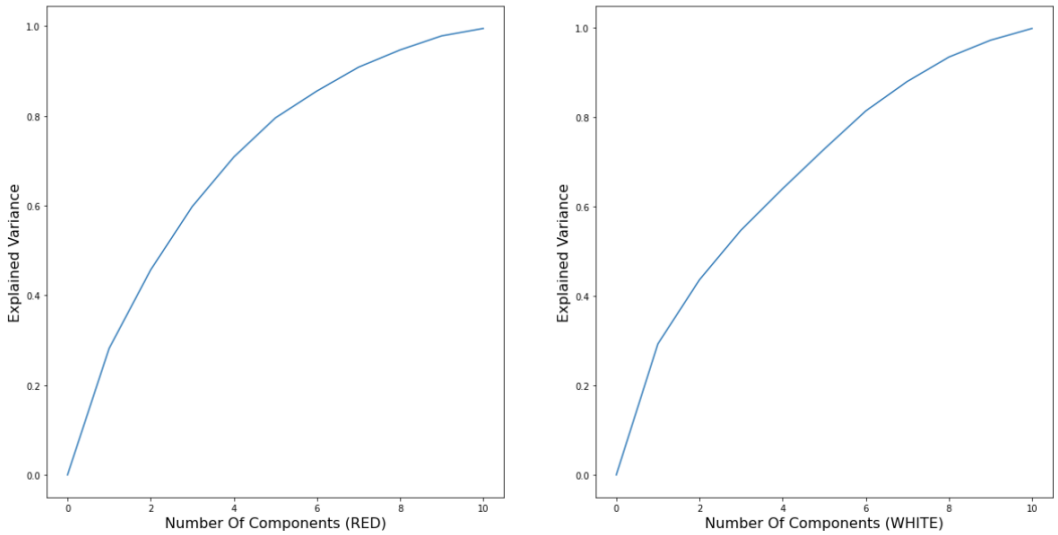

Figure 3: Coefficients for Major Factors in Wine