Analysis of wine dataset

170183149

November 24, 2020

0.1 Introduction

The wine dataset that will be analysed in this report is easy to find and download from the internet by searching 'wine dataset'. Firstly, the imported libraries are pandas, numpy, matplotlib and sklearn. The datasets are imported using the pandas command `pd.read_csv()`, under the names 'all' - the combined file, and 'red' and 'white' - the files separated by colour. Before the data can be analysed it has to be cleaned in order to provide more accurate results. There were no missing values in this dataset as was checked by viewing the number of rows in 'all' using `all.shape` and matching them to the count of every column using `all.describe()`. Duplicate observations skew the data so they were removed from each dataset using `drop_duplicates(inplace=True)`, the argument there saving the changes to the dataset. Using `sns.pairplot(all)` to create scatterplots of all pairwise components, anomalies can be spotted by checking the row or columns of each variable for extreme outliers. Taking note of the approximate place of the anomaly, the exact location is found by `all.loc[all["residual sugar"]>60]`, in this example the variable displaying the anomaly is 'residual sugar' and the value it was observed to be outside of is 60. The anomalous observations were removed from the combined dataset and the dataset of whichever colour it was with the command `drop([], axis=0, inplace=True)` with the index number of the observation written inside the square brackets. Note that the index number for rows in the 'red' data corresponds to the index number of the same observation in the combined file minus 4898 which is the number of white wine observations. Overall, five observations were removed. No columns displaying extra variables were added at this stage as a column being a linear combination of two other columns would have complicated the Principal Component Analysis by adding a 0 eigenvalue.

0.2 PCA

To begin the analysis, PCA (Principal Component Analysis) is performed on the combined file. The correlation matrix was used for this process as some of the variables have much higher variance than others for example total sulfur dioxide at 56.774 compared to density at 0.008. To begin with the data was

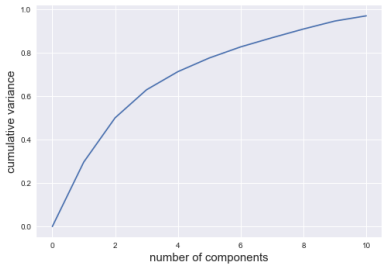

Figure 1: Cumulative variance

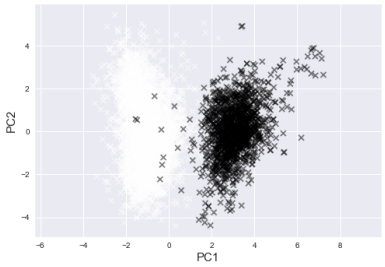

Figure 2: PC1 vs PC2

scaled so each variable had unit variance,

```
scaler=StandardScaler()
scaler.fit(all)
all_sc=scaler.transform(all),
```

the second line calculating the mean and variance and the third fitting it to the data. PCA is applied using this code,

```
pca=PCA()
pca.fit(all_sc),
```

and the cumulative totals of the proportion of variance taken up by each component can be found using `np.cumsum(pca.explained_variance_ratio_)` and are shown below.

```
[ 0.2964303   0.50025232  0.62913351  0.71263252  0.77552477  0.82657166
  0.868662    0.9087577   0.94544147  0.96944433  0.98915287  0.99833571]
```

From this and the Figure 1 scree plot it can be seen that after the tenth component only small proportions of variance are added from the extra components therefore we can add the argument `n_components=10` to `pca=PCA()` which allows us to reduce the dimensions from 12 to 10. The function `pca.components_` produces an array of the weightings for each variable. After fitting the weightings to the data using `p=pca.fit_transform(all_sc)` we plot PC1 against PC2 (Figure 2) and colouring the white wine data white and the red wine black we can see that there are two distinct classes which hints that there are major chemical differences between red and white wine. The wines perform similarly on PC2 but the red wines do better on PC1. Looking at the weightings for PC1 compared to the index,

```
[  2.68362475e-01   3.61749862e-01  -1.09546477e-01  -2.11453546e-01
    3.09651641e-01  -3.32870844e-01  -3.88927273e-01   2.10114617e-01
    1.57890984e-01   2.82826258e-01  -4.33558282e-02  -1.07421711e-01
    4.69005828e-01]
['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
      'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
```

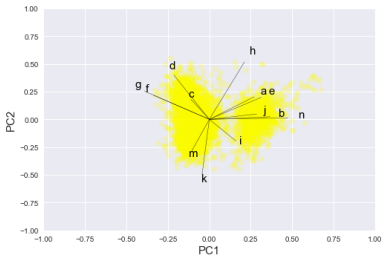

Figure 3: PC1 vs PC2

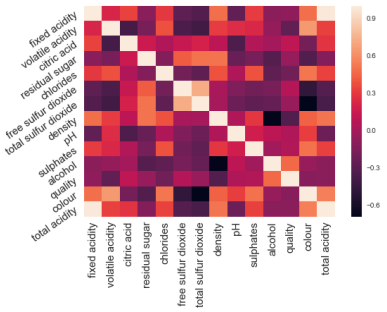

Figure 4: Heatmap of all wines

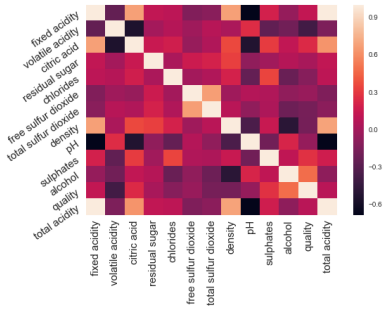

Figure 5: Heatmap of red wines

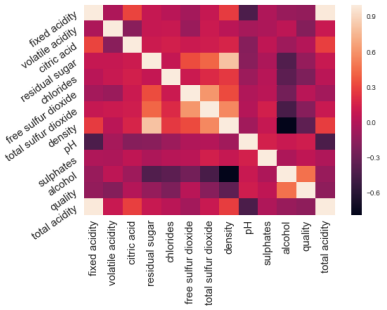

Figure 6: Heatmap of white wines

```
'pH', 'sulphates', 'alcohol', 'quality', 'colour'],
```

for red wine to have a higher PC1 it is more likely to have higher amounts of fixed acidity, volatile acidity, chlorides, density, sulphates and pH compared to white wine and lower amounts of citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide and alcohol. Plotting every other principal component against each other, no other distinction between the classes stood out as much as the differences in PC1. The biplot in Figure 3 illustrates how each variable affects PC1 and PC2 and agrees with the analysis of the coefficients (labels a-n corresponding to column names).

0.3 Quality

Investigating the effects of different variables on quality, firstly the wines with the highest quality rating were located using `max(all.quality)` (which was 9) and then `all.loc[all["quality"]>8]` to find all the observations that received a 9. Comparing each variable to the mean for its type, only one feature was present in all 5 top wines which was that their chloride levels were all less than the average of 0.056. All 5 were white wines. Total acidity might be easier to work with than both fixed and volatile acidity so that is added to the

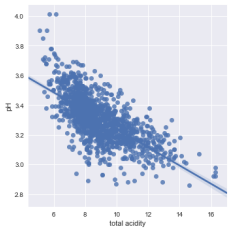

Figure 7: pH vs Total Acidity - Red

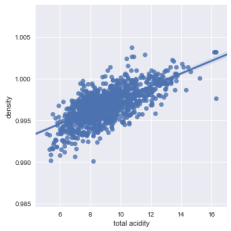

Figure 8: Density vs Total Acidity - Red

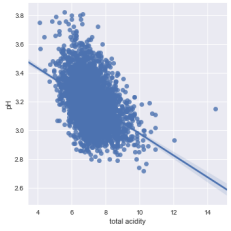

Figure 9: pH vs Total Acidity - White

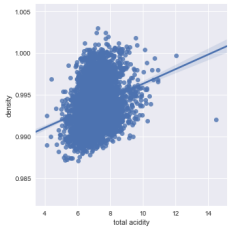

Figure 10: Density vs Total Acidity - White

data. Looking on the heatmap down the 'quality' row the only strongly positive correlation that stands out is with alcohol. It would be hard to model quality based on the other variables as it is a subjective quantity and only takes a very limited range of values. Meaning that although there could be a lot of variation in the input there would not be much or any in the output. This is reflected by the neutral colours in the quality row of the heatmap. Comparing the red and white heatmaps, the red dataset shows much stronger correlation in general between its variables for example for red wines, total acidity has a large effect on density and pH but for white wines they don't have as much of a connection.

0.4 Conclusion

From examining the datasets we can see that red and white wines are very different chemically and PCA has produced an idea of what the main chemical differences could be. There are factors that are correlated to good quality wine however nothing strong enough to suggest a model could be made and it will always be down to personal preference as is shown by the top 5 wines all being white. To confirm the accuracy of this, further analysis would need to be made to predict the outcome of an observation then compare with the real ratings.