Sushidhar Jayaraman
1001400523

This is the analysis on wine data quality from kaggle. The data set has 4898 observations with the class variable 'quality' on the scale of 1 – 10 range.

Normalization:

Original Class variables distribution:

	quality	count
0	6	2198
1	3	20
2	5	1457
3	9	5
4	4	163
5	8	175
6	7	880

The extreme low and high class has very less count compared to the class in the middle.
So the class has been reduced to the range "Low", "Medium", "High".

Undersampling:

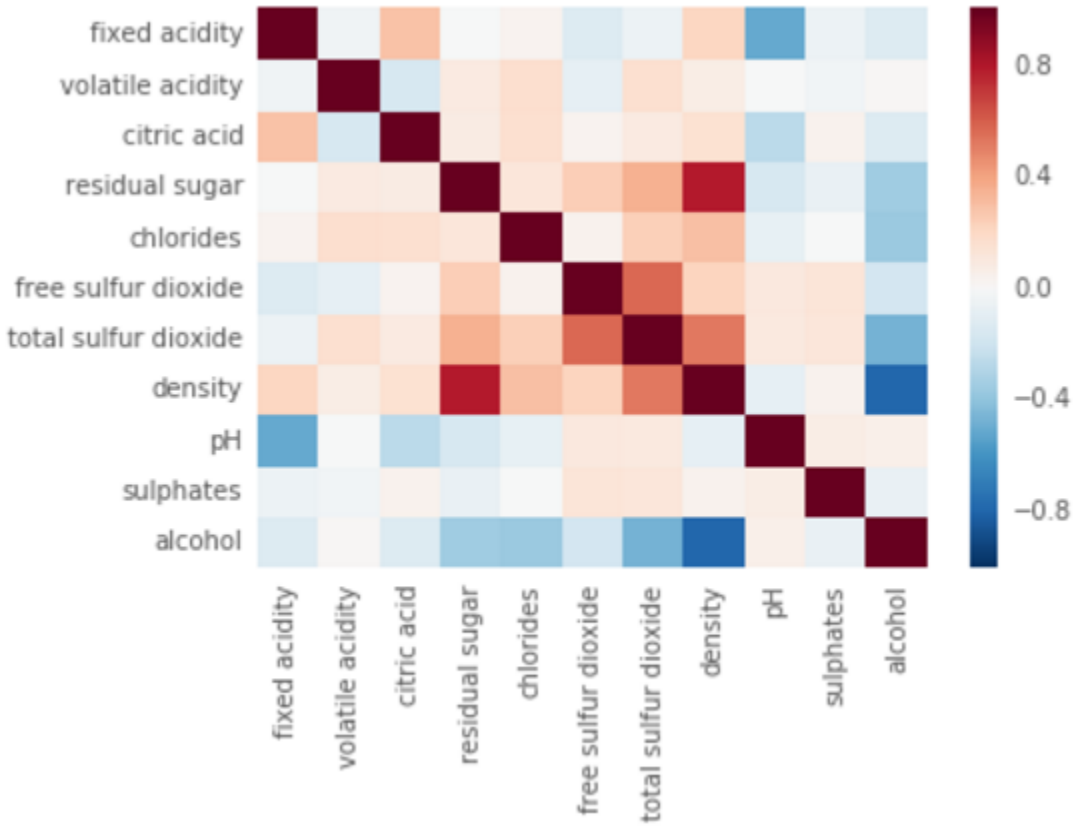
There is an imbalance in the dataset after the class has been reduced to "Low", "Medium", "High".
To deal with it stratified sampling has been done the dataset. After that the distribution has been reduced to:

	quality	count
0	High	1060
1	Low	1078
2	Medium	1011

Sushidhar Jayaraman
1001400523

Feature Selection:

Following heatmap is created to find the correlation between the variables



From the chart, it is evident that there is a strong correlation between the density and residual sugar as well as between the total sulfur dioxide and free sulfur dioxide. So, one feature has been removed between the pairs and accuracy, f1-score has been calculated to find out which is the best feature in the pairs. So, we found out that there is high accuracy when dropping the feature free sulfur dioxide and density.

Feature	**Decision Tree - Accuracy**
total sulfur dioxide and density	0.5663590
free sulfur dioxide and density	0.5950792
total sulfur dioxide and residual sugar	0.5724541
free sulfur dioxide and residual sugar	0.5854921

Sushidhar Jayaraman
1001400523

Principal Component Analysis:

Following table shows the pcaFeatures (by weighted sums) to find orthogonal directions of maximum variance. It has a tuple of (eigenvectors, `RDD` of scores, eigenvalues). Eigenvectors is a multi-dimensional array where the number of rows equals the length of the arrays in the input `RDD` and the number of columns equals 3. The `RDD` of scores has the same number of rows as `data` and consists of arrays of length 3. Eigenvalues is an array of length d (the number of features).

```
+-----------------------------------------------------------------+
|pcaFeatures                                                      |
+-----------------------------------------------------------------+
|[-176.52109231240544,0.953114906178876,-11.480159664959078]      |
|[-130.86253868448225,21.148918749176485,5.152966126208715]       |
|[-101.56366694584099,-3.5484737726864894,-1.1708994508110688]    |
|[-191.87569176299675,3.4045071682030277,1.4883166424660763]      |
|[-191.87569176299675,3.4045071682030277,1.4883166424660763]      |
|[-101.56366694584099,-3.5484737726864894,-1.1708994508110688]    |
|[-139.16209994016236,6.682688895277058,0.38126217006329155]      |
|[-176.52109231240544,0.953114906178876,-11.480159664959078]      |
|[-130.86253868448225,21.148918749176485,5.152966126208715]       |
|[-131.61653791341118,6.863264429987194,5.593801454673983]        |
|[-63.535192022347815,5.916684491874887,2.4521410215420407]       |
|[-109.6035672652093,12.192737853893329,1.6463318013292367]       |
|[-76.41372887119383,4.255649503959447,3.251299462210956]         |
|[-150.32826755633715,-8.77495941700684,6.858360674878846]        |
|[-177.32361682238408,5.365374155012407,-9.99958158489786]        |
|[-115.22364755809386,2.3803767412359154,4.919367335358098]       |
|[-103.22023857592676,-2.949266752305727,4.668606209457103]       |
|[-79.79053280916112,-8.313103488034857,3.7817429724780087]       |
|[-169.19419104428871,28.50758847663812,7.531902492510495]        |
|[-137.34341432527427,2.031659620982571,-0.15089050076385757]     |
+-----------------------------------------------------------------+
only showing top 20 rows
```

Sushidhar Jayaraman
1001400523

Algorithm:

Four Model is built on the wine quality dataset out of which 3 classification algorithms and 3 regression algorithms.

1. Linear Regression

2. Decision Tree Classification and Regression

3. Random Forest Classification and Regression

4. Support Vector Machine Classification. – Since SVM is a binary classifier OneVsRest classifier is used which supports multiclass values.

Classification Algorithm:

Algorithm	Accuracy	F1-Score
Decision Tree without Normalization	0.579587628866	0.556737226561
Decision Tree with Normalization	0.655079242454	0.621779242175
Random Forest without Normalization	0.569531841958	0.539147670961
Random Forest with Normalization	0.634561154012	0.624691358025
SVM – OneVsRest without Normalization	0.567110157368	0.547018102814
SVM – OneVsRest with Normalization	0.648109820486	0.622325941107

Regression Algorithm:

Algorithm	RMSE
Decision Tree without Normalization	0.794942090862
Decision Tree with Normalization	0.776409755839
Random Forest without Normalization	0.753327043668
Random Forest with Normalization	0.642378711939
Linear Regression without Normalization	1.05893575031
Linear Regression with Normalization	0.782533132629