Abishek Ganesh (1001443953)

Big Data Machine Learning Homework

Data Description:

Dataset is for white wines is taken from UCI, http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Given data has the following attributes:

Features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

Target Variable: quality (score between 0 and 10)

There are no missing values in the given data set

Total no of features: 11

Total no of Rows: 4898,

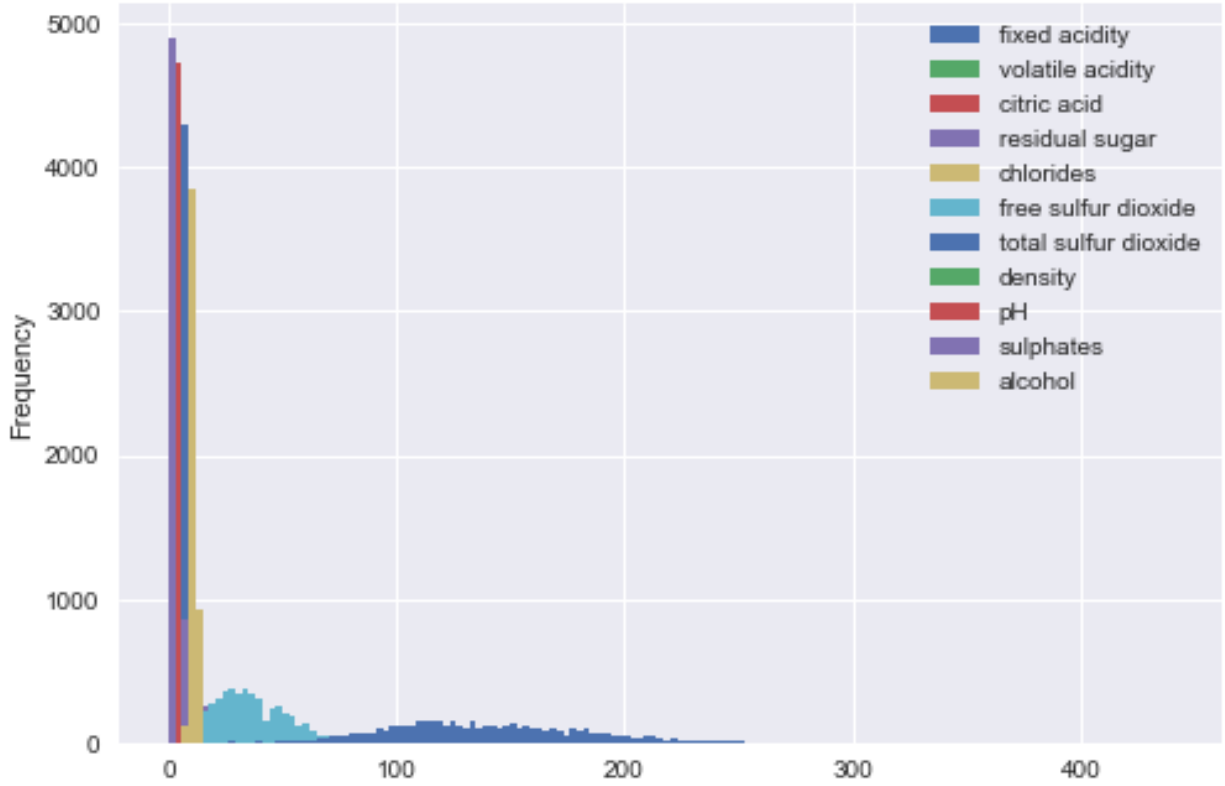
Sample of Data:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
1	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
2	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.3	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.2	0.4	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.2	0.4	9.9	6

Data description:

	count	mean	Std	min	25%	50%	75%	max
fixed acidity	4898	6.854788	0.843868	3.8	6.3	6.8	7.3	14.2
volatile acidity	4898	0.278241	0.100795	0.08	0.21	0.26	0.32	1.1
citric acid	4898	0.334192	0.12102	0	0.27	0.32	0.39	1.66
residual sugar	4898	6.391415	5.072058	0.6	1.7	5.2	9.9	65.8
chlorides	4898	0.045772	0.021848	0.009	0.036	0.043	0.05	0.346
free sulfur dioxide	4898	35.308085	17.007137	2	23	34	46	289
total sulfur dioxide	4898	138.360657	42.498065	9	108	134	167	440
Density	4898	0.994027	0.002991	0.98711	0.991723	0.99374	0.9961	1.03898
pH	4898	3.188267	0.151001	2.72	3.09	3.18	3.28	3.82
sulphates	4898	0.489847	0.114126	0.22	0.41	0.47	0.55	1.08
Alcohol	4898	10.514267	1.230621	8	9.5	10.4	11.4	14.2
Quality	4898	5.877909	0.885639	3	5	6	6	9

Graphical Representation of Data:



Data Cleaning

Class Imbalance

The given data is highly imbalanced:

Class Labels	**Count of quality**
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5
Grand Total	**4898**

So, we correct the imbalance by splitting the data to 3 labels

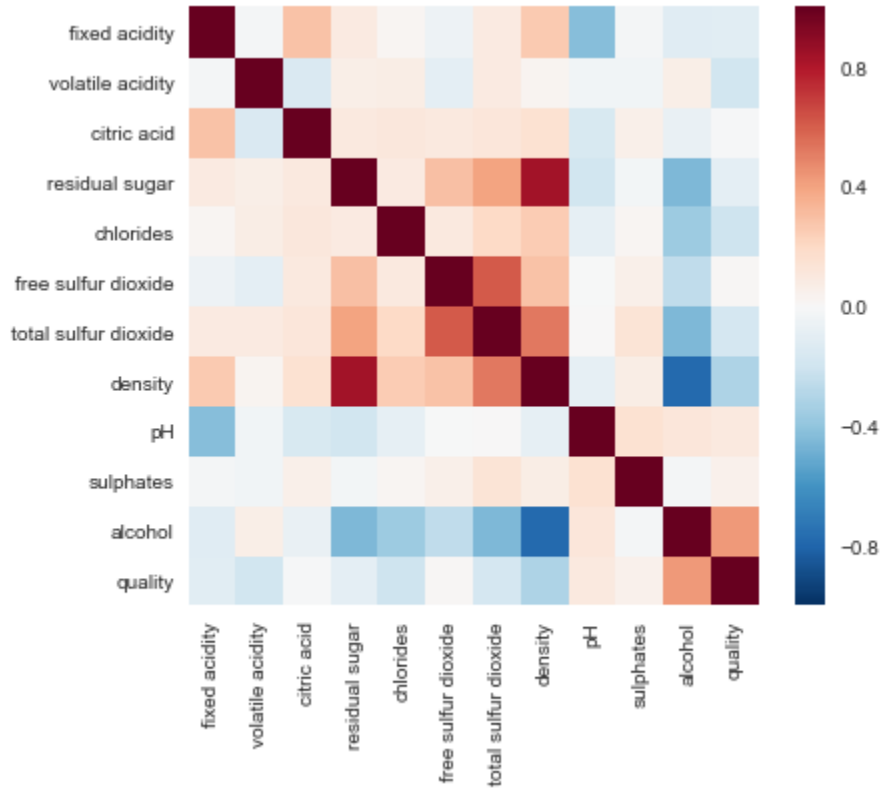
- Low: represented by ‘0’, having 3, 4, 5 labels
- Medium: represented by ‘1’, having 6 label

- High: represented by ‘2’, having 7, 8, 9 labels

Correlation

Since the given data is numeric so we will check the correlation between the features

Correlation Graph:



After seeing the heat map for correlation, it is clearly visible that

- Density and alcohol
- Residual sugar and Density
- Total sulfur dioxide and Free Sulfur Dioxide

Are highly correlated and it is not very useful to have highly correlated values in the dataset so removing features:

Density and Free sulfur dioxide

Outliers

Data contained outliers in fixed acidity, volatile acidity, citric acid, residual sugar, chlorides and they were removed.

Steps Followed

- Balanced the target variable as the classes were highly skewed by grouping the various data into groups and relabeling them as low, medium, high.
- Checked for correlation between the features. Identified 3 pairs of highly correlated variable and removed 2 features which were: Density and Free Sulfur Dioxide
- Checked for outliers by plotting the and removed them to prevent bias.
- Ran the given algorithms with cross validation and 70-30 split.

Performed the following operations:

- Class Balancing
- Feature removal
- Normalization
- used Std Scalar function
- PCA
- Sampling

Algorithms

the algorithms that are being used are the following:

a) Logistic Regression
b) Decision Tree (as Classifier)
c) Random Forest (as Classifier)
d) SVM (using One vs Rest approach)

Not using Linear Regressing as the class label is nominal

All the algorithms are used 70-30 split

Summary of Runs:

Original		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.45708155	0.26837496
Random Forest	0.53648069	0.49638085
Decision Tree	0.43204578	0.3582671
SVM	0.53576538	0.49931824

Class Balaced		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	**0.45815451**	0.28676986
Random Forest	0.57796853	0.57084327
Decision Tree	0.5944206	0.59345141
SVM	0.56366237	0.55540974

Feature Removed		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.56008584	0.53275927
Random Forest	0.57939914	0.57111206
Decision Tree	0.5851216	0.58437147
SVM	0.5658083	0.55665389

Normalization of data		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.49928469	0.44339266
Random Forest	0.57296137	0.5336279
Decision Tree	0.55293276	0.53866217
SVM	0.57796853	0.56531976

Standardization of data		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.51716738	0.51085441
Random Forest	0.57796853	0.57084327
Decision Tree	0.58520601	0.59345141
SVM	0.56938484	0.56253066

With Sampling		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.4795082	0.3878467
Random Forest	0.57991803	0.57861787
Decision Tree	0.55122951	0.54712298
SVM	0.56762295	0.56648892

Outliers Removed		
Algo	**Accuracy**	**F1 Score**
Logistics Regression	0.48284642	0.43392659
Random Forest	0.56488011	0.55340618
Decision Tree	0.59155148	0.59435051
SVM	0.56981664	0.56083061

Conclusion

- After Balancing the class, we can see that the accuracy and F1 Score has improved for all the classifiers.
- For Classifiers removing outliers led to a marginal decrease in the accuracy and F1 score, this can be accounted to the fact that the outliers were unique values.
- After doing normalization and we can see a clear improvement in the accuracy and F1 score
- After doing sampling we can see a reduction in accuracy and F1 for Decision Tree Classifier which can be because of the sample selection, some unique values might have been lost.