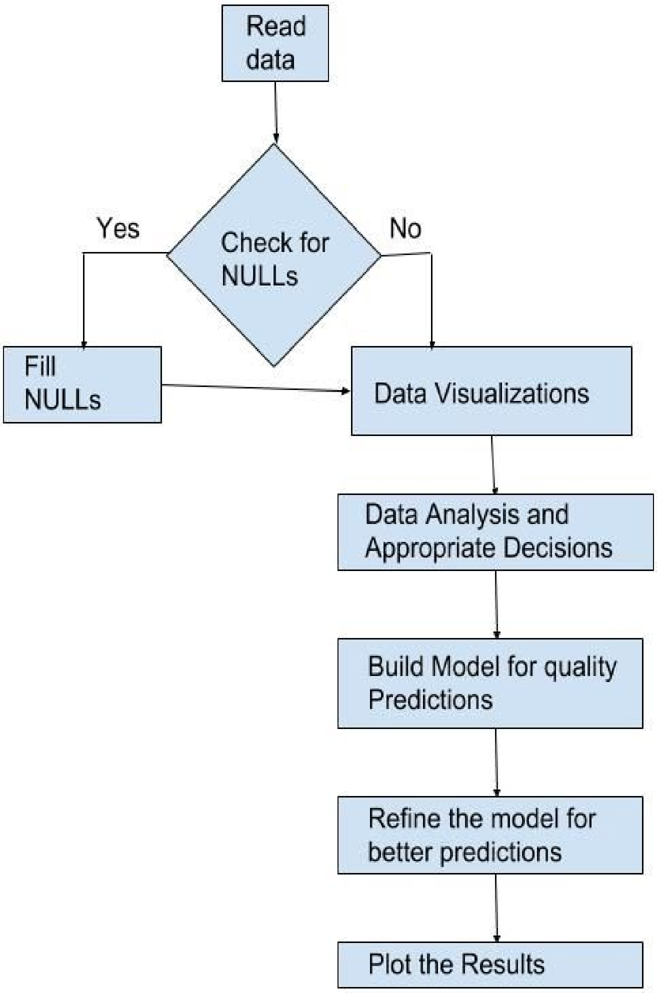
1. Problem Definition:

Given the Portugal wine dataset, the aim is to analyze data and build a model that could better approximate the quality of unseen test data set.

2. Abstract:

The aim of the project is to report various data visualization techniques and data modeling for prediction as part of data analysis of the famous wine dataset of Portugal. All the implementation has been done in Python using various libraries such as Pandas, Seaborn, Matplotlib, sklearn, etc. The goal of this analysis is to extract the most influencing factors on the wine quality in the given wine dataset. In this report, we could see the approach to achieve the above-mentioned task.

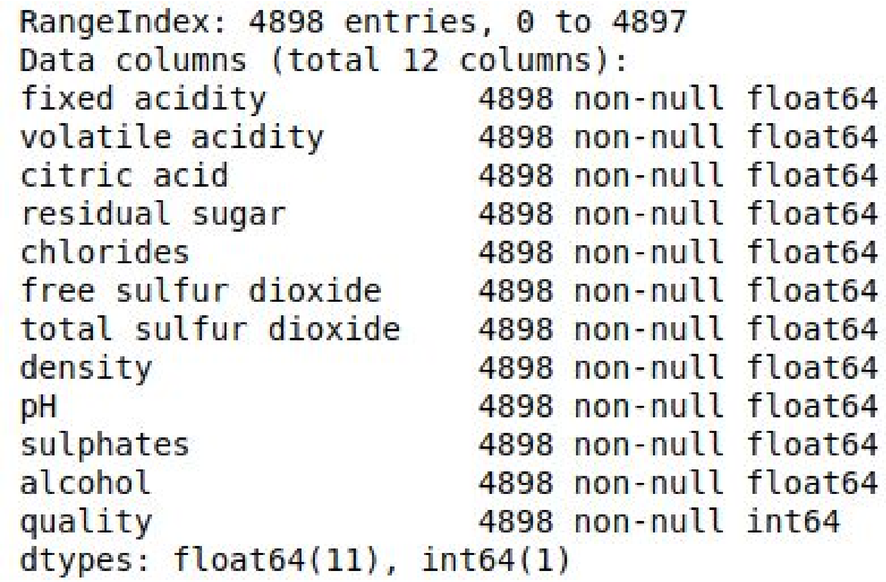
3. Methodology:


Read data
Check for NULLs
Yes
No
Fill NULLs
Data Visualizations
Data Analysis and Appropriate Decisions
Build Model for quality Predictions
Refine the model for better predictions
Plot the Results

4. Results:

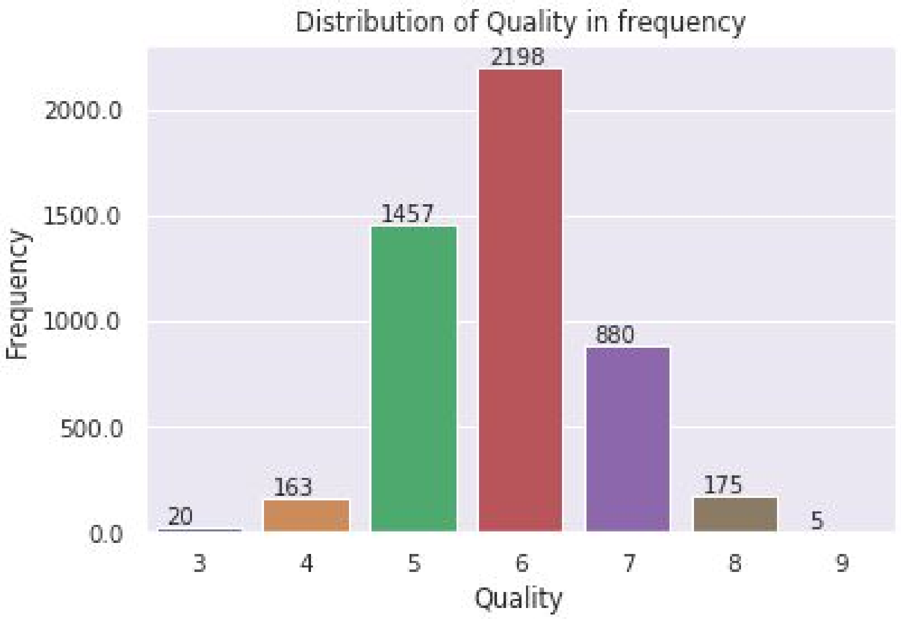
I have basically experimented using 4 algorithms namely linear regression, KNN, decision tree, and neural networks. Linear regression is used to predict the quality of the test wine dataset whereas the decision tree, KNN and neural network were used to predict the class to which the quality belongs to, namely high, medium, low. Further details are presented in the Discussions section. Upon applying linear regression, the R-squared score turned out to be around 0.26. The accuracy of the Decision tree was around 0.586. KNN gave an accuracy of around 0.77 and that given by the neural network was around 0.715.

5. Discussion:

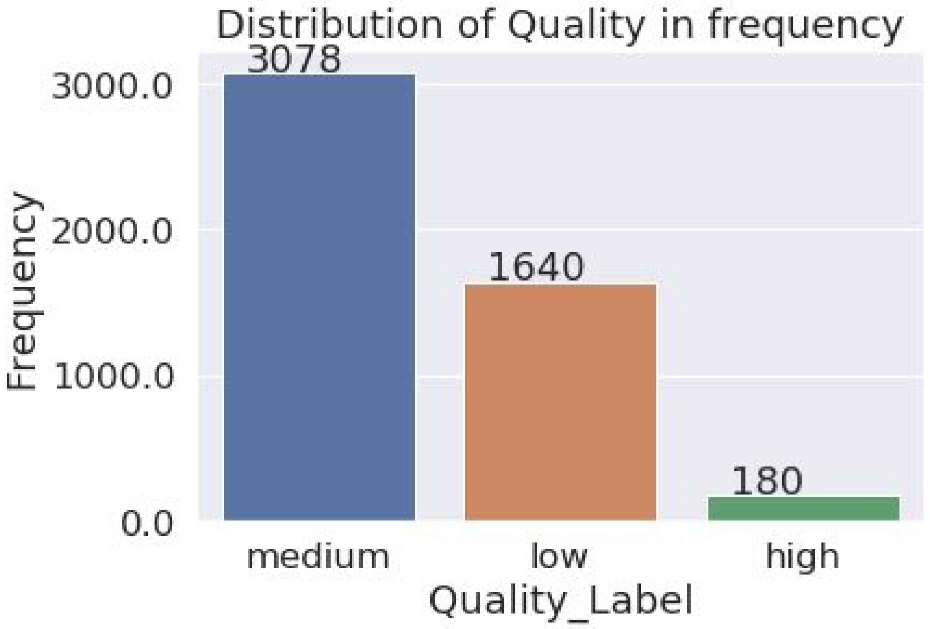


```
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
```

The data has 4898 entries and 12 columns in all. One can see the features of the data represented by the columns. As evident from the picture, there are no null values in the data. Hence, let us proceed to the data visualization part. The frequency of different types of quality is distributed as follows.

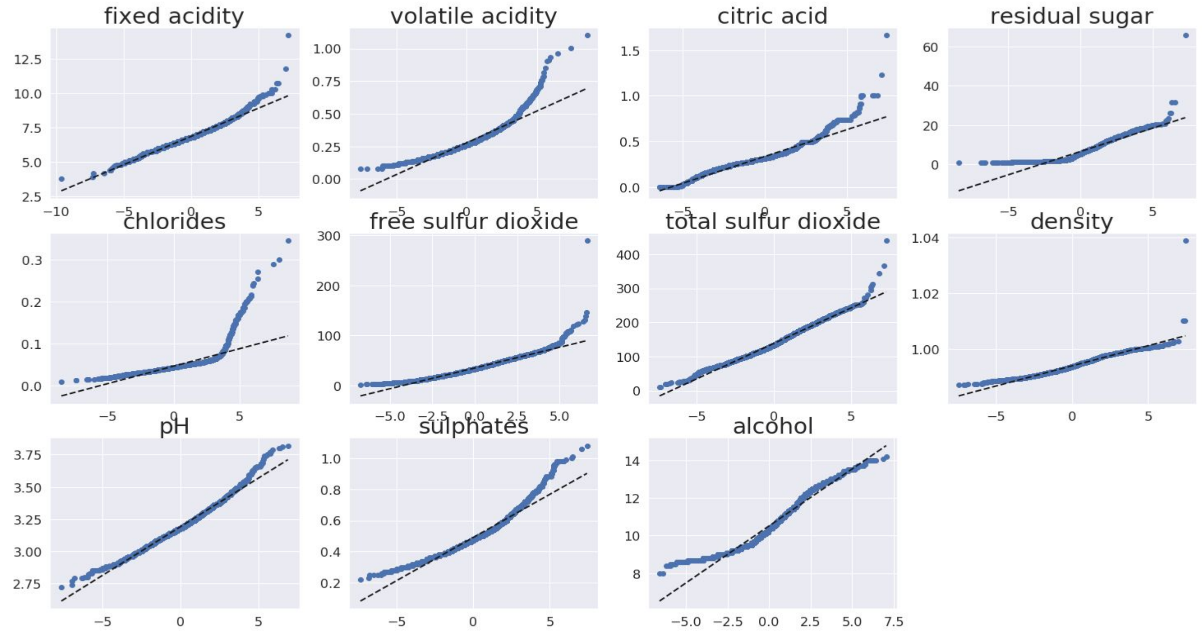


The data has also been divided into three classes namely those having quality low, medium and high. If the quality is less than or equal to 5, it is labeled to have low quality, if the value is above 5 and less than or equal to 7 then it is medium and above 7 is treated to be a high-quality one.

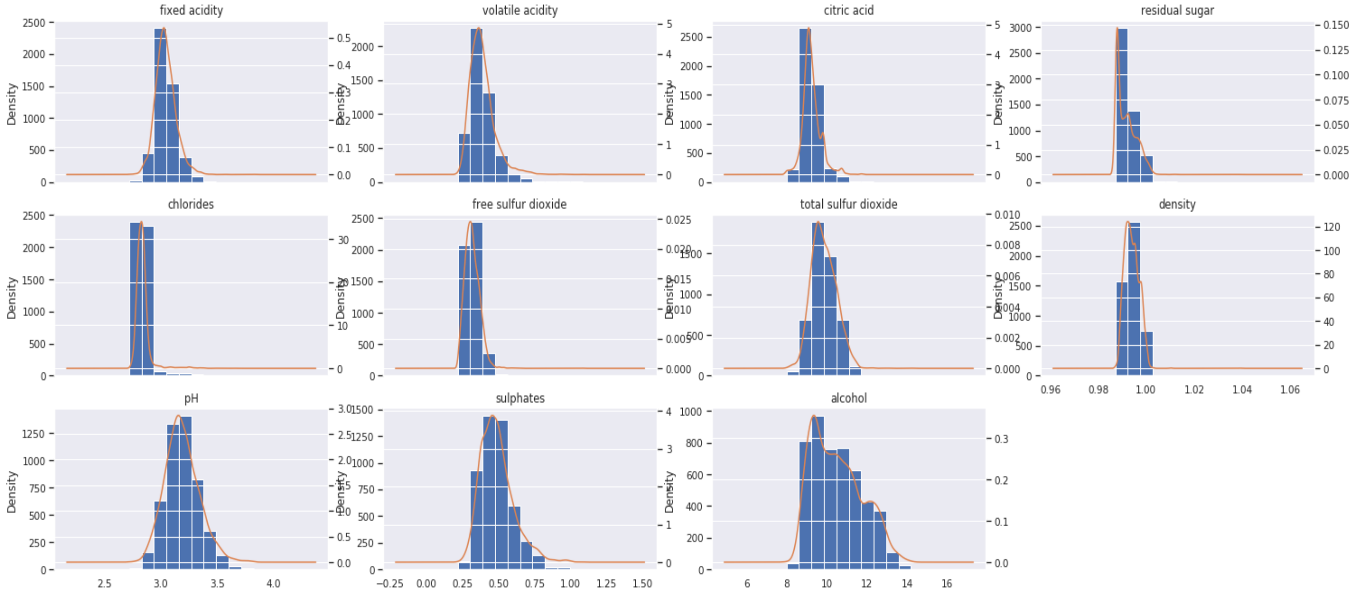


Q-Q Plots

Many standard statistical procedures require normally distributed data. One way to assess if your data is normally distributed is the quantile-quantile plot or q-q plot. In this approach quantiles of a tested distribution are plotted against quantiles of a known distribution as a scatter plot. If distributions are similar the plot will be close to a straight line. We will plot our data against a normal distribution to test if our data is distributed normally. The distributions of the total sulfur dioxide, density, and fixed density almost approached the normal distribution.

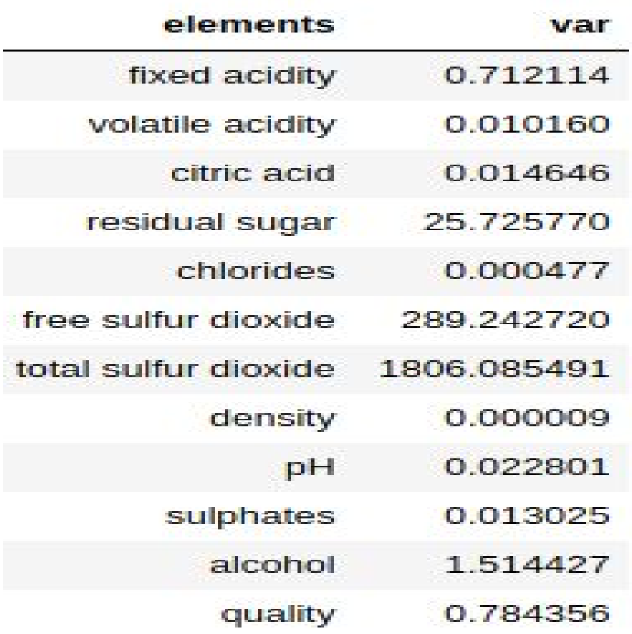


Distribuation Plots



Below is a table showing the variance of the distribution of each feature.

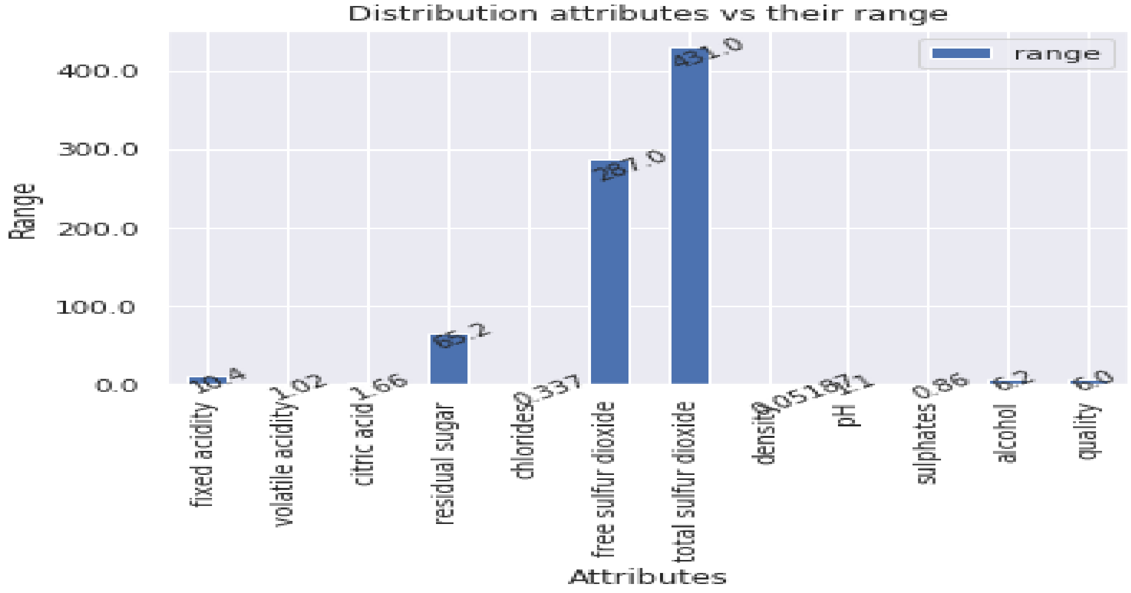
Attributes vs Variance



elements	var
fixed acidity	0.712114
volatile acidity	0.010160
citric acid	0.014646
residual sugar	25.725770
chlorides	0.000477
free sulfur dioxide	289.242720
total sulfur dioxide	1806.085491
density	0.000009
pH	0.022801
sulphates	0.013025
alcohol	1.514427
quality	0.784356

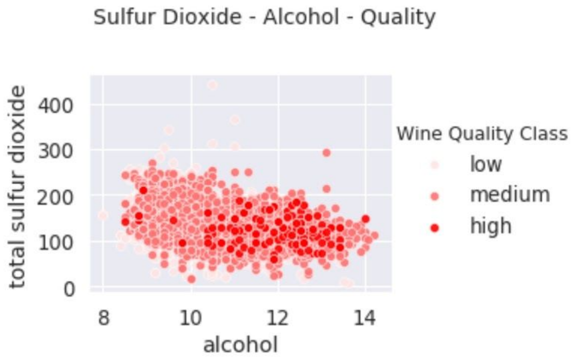
Range of values

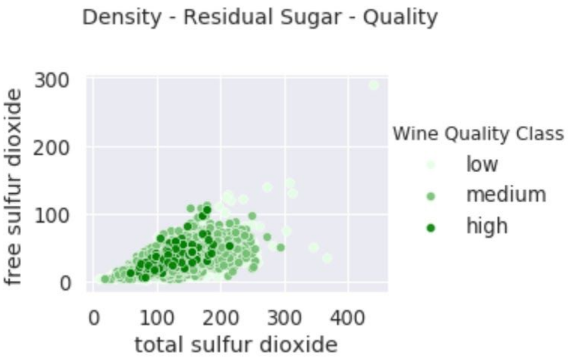
The ranges of the attributes are as follows.

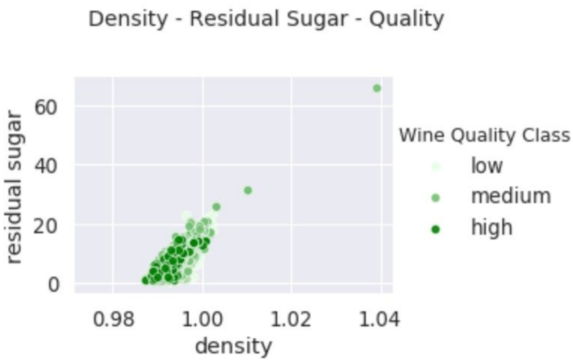


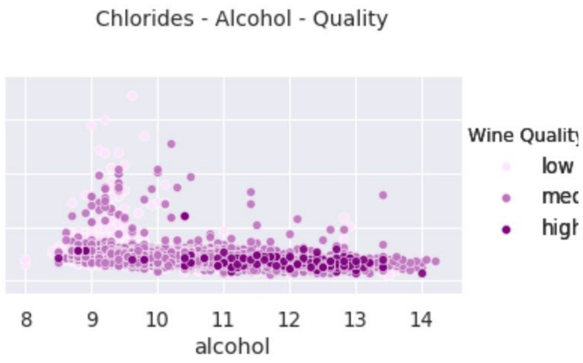
Attribute-Correlation Plots

It is always interesting and often important to find if the attributes are correlated to each other. The heat map shows how each of them are correlated with each other.
The better-correlated attributes with the quality classes are shown below.

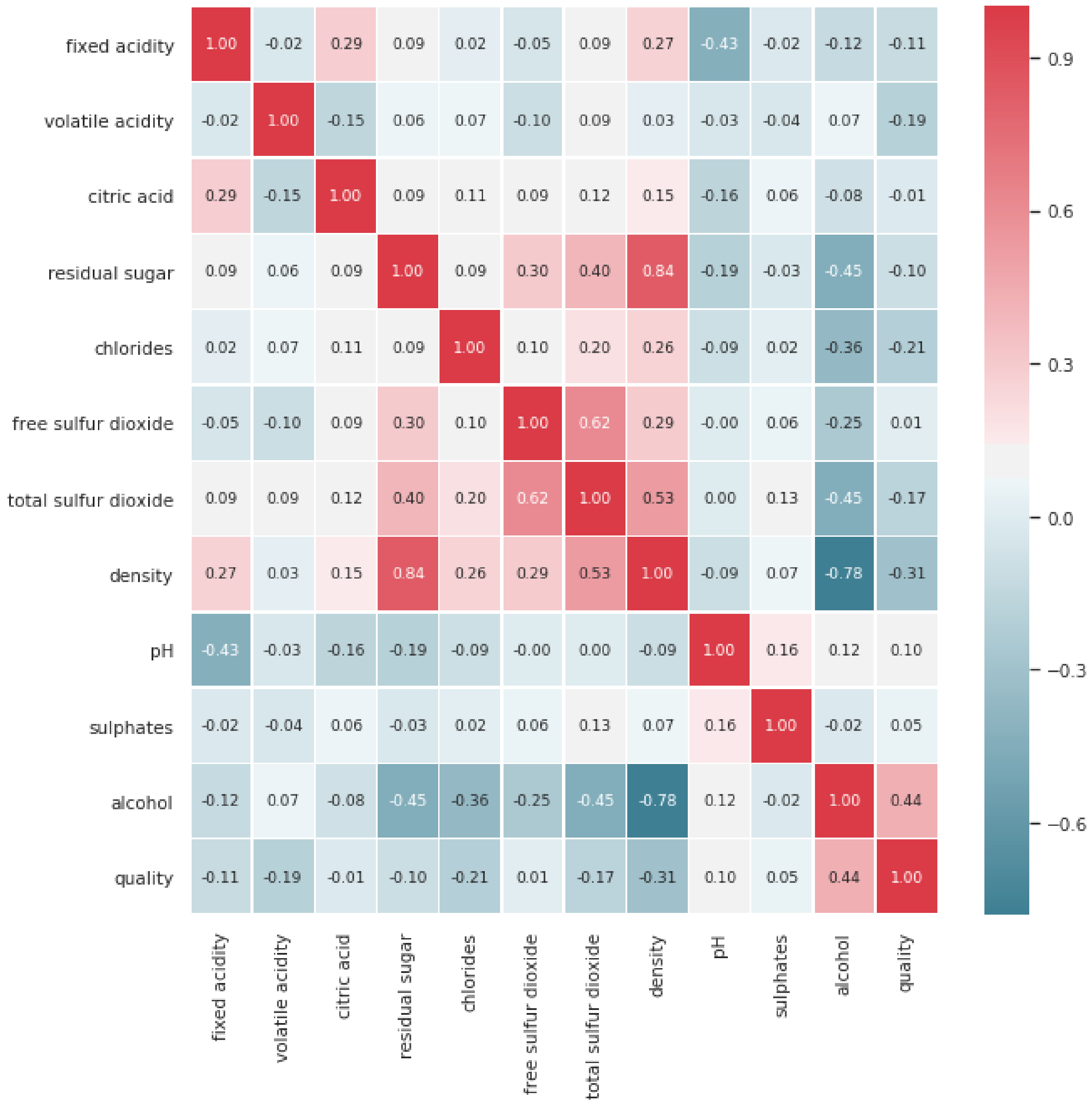

Sulfur Dioxide - Alcohol - Quality


Density - Residual Sugar - Quality


Density - Residual Sugar - Quality


Chlorides - Alcohol - Quality

The Heat map



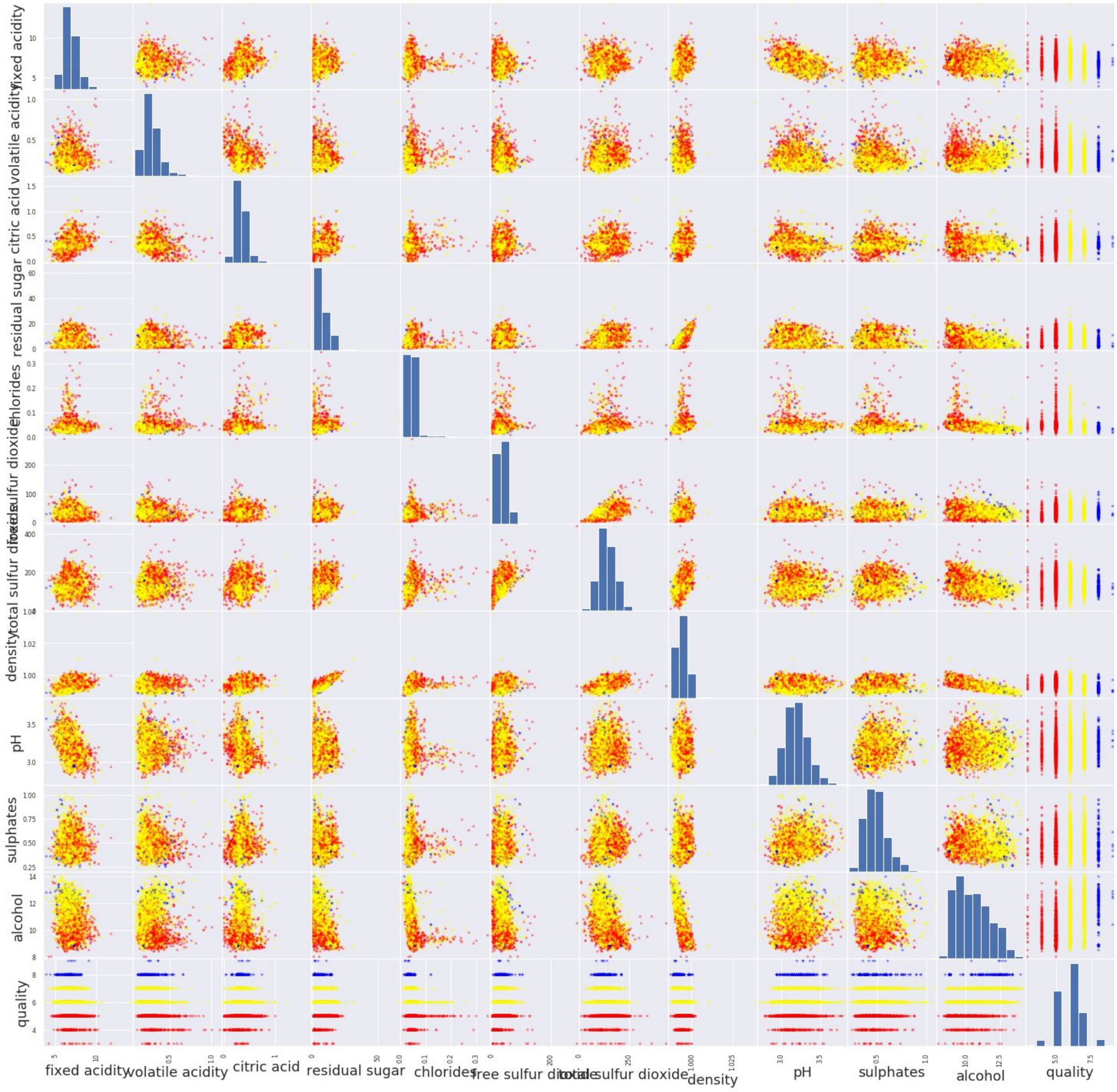
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07	-0.19
citric acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08	-0.01
residual sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45	-0.10
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36	-0.21
free sulfur dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	-0.00	0.06	-0.25	0.01
total sulfur dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45	-0.17
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78	-0.31
pH	-0.43	-0.03	-0.16	-0.19	-0.09	-0.00	0.00	-0.09	1.00	0.16	0.12	0.10
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02	0.05
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00	0.44
quality	-0.11	-0.19	-0.01	-0.10	-0.21	0.01	-0.17	-0.31	0.10	0.05	0.44	1.00

Inferences from the heatmap:

- Total and free sulfur dioxide has the highest correlation.
- Density has a relatively high positive correlation to residual sugar and relatively high negative correlation to alcohol.
- The residual sugar has a subtle positive correlation with the total sulfur dioxide and with free.
- Chlorides and alcohols have a negative correlation with each other.
- Chloride and residual sugars have a negative correlation with each other. Also, alcohol and total sulfur dioxide are negatively correlated with each other.

The following depicts the correlation of attributes with the quality classes high, low and medium.

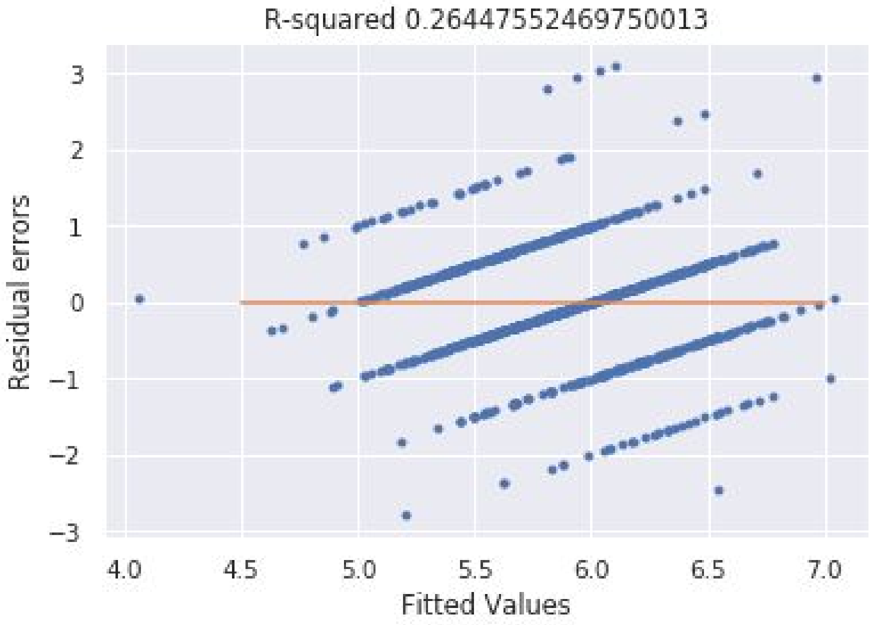
Scatterplot for attribute vs quality label



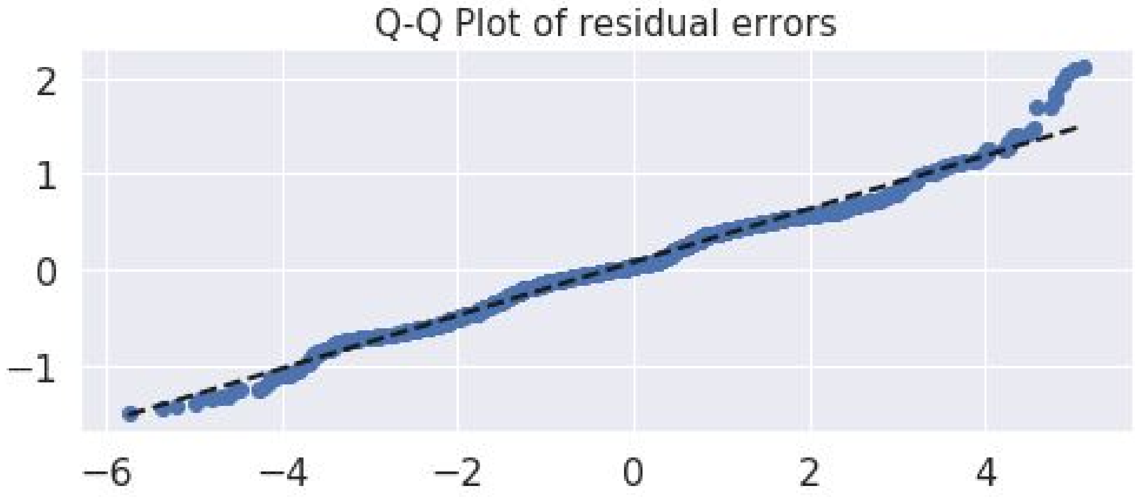
6. Modeling:

6.1 Linear Regression:

Linear regression has been implemented using sklearn, a machine learning library for Python. The data was split into the train and the test parts in the ratio 80:20. Considering the raw attributes for modeling, the R-squared value turned out to be around 0.26. The mean square error was around 0.68. The plot of residual errors versus the fitted values followed heteroscedasticity. Then, I transformed some data values using the Box-Cox method. After the processed values were used for training, now the MSE turned out to be around 0.3.



The following graph depicts that the residual errors nearly follows the normal distribution.

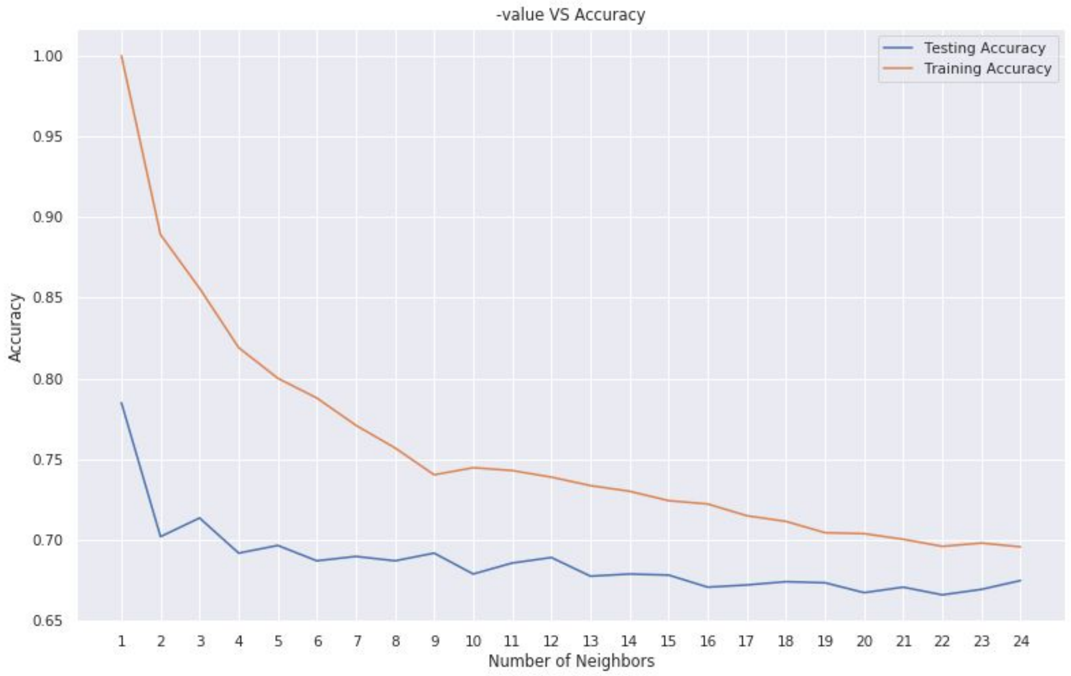


6.2 Decision Tree:

Next model that was built was the decision tree. Deciosion tree is an unsupervised algorithm built upon the criteria such as information gain, and Decision each level intending to finally come up with a model for accomplishing the classification problem. Considering the class of the quality to be the target value to be found out, the decision tree upon getting trained on 80% of the data and tested upon 20% of the data gave an accuracy of around 0.59. The MSE was 0.79. For reference, the image of the tree constructed was attached separately.

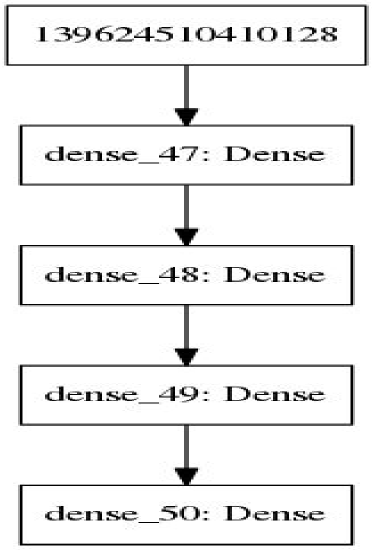
6.3 KNN:

KNN is yet another unsupervised classification algorithm. It uses K nearest neighbors to classify the new incoming data point. On this given data, this algorithm has achieved the best accuracy of around 0.785 with k=1. The following is the plot of k-value versus accuracy.



6.4 Neural Network:

The last algorithm used was the dense neural network. The model built was as follows. Using Adam optimizer at a learning rate of 0.001, it has achieved an accuracy of 0.715 with an MSE of 0.13.



7. Conclusions:

The results of the models built on the real world data are quite different than what is generally studied. Analysis of the nature of the data and its appropriate transformation are often essential in order to refine the model we are building. The linear regression, after applying such data transformations has actually given a lower mean squared error than that applied on the raw data for prediction of the quality. On the other hand, considering the classification problem of the given wine features into 3 broad classes of high, medium and low, KNN with the k value to be 1 with an accuracy of around 0.77 has out-performed the decision tree classifier which gave an accuracy of around 0.593. Nevertheless, the neural network also stands as an equivalent competitor to the KNN with an accuracy of around 0.715 also with as least as 0.13 MSE, the lowest among all. Since the dimensionality of the data is not so big and since the features also mostly not so well correlated, pruning off of the features with PCA or removing correlated data was not considered. The result may be more promising if more brilliant and contributing features were considered while collecting the dataset.