Red Wine Quality Prediction



Rasha Jamsheer

Data Trained Education, Noida

Problem Statement

The dataset is related to red and white variants of the Portuguese “ Vinho Verde" (literally ‘green wine’)wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (E g : there is no data about grape types, wine brand, wine selling price, etc.).  
  
This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

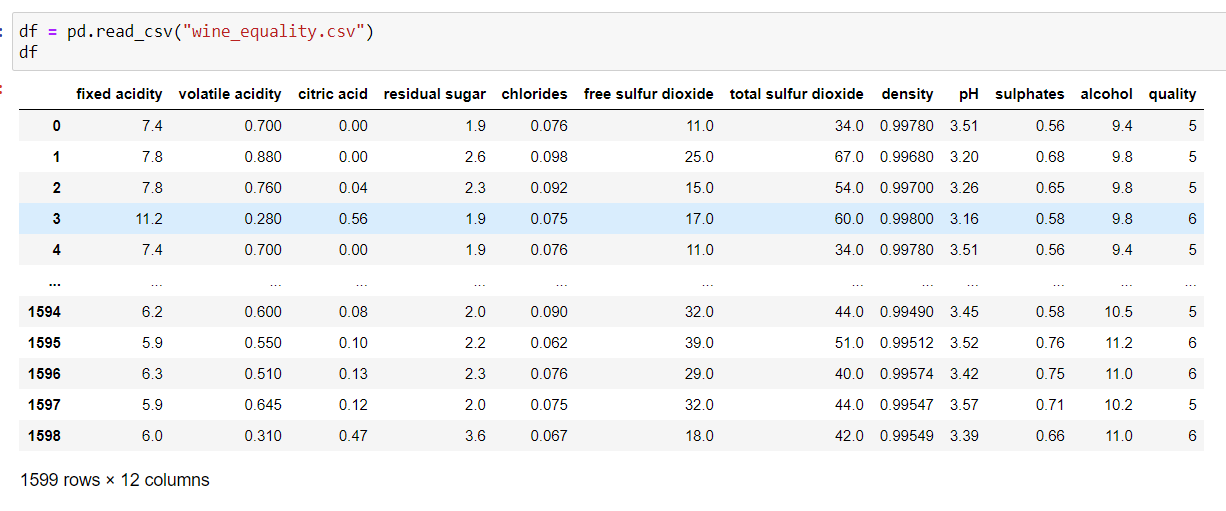
## Description of Dataset

When we download the dataset , we can see that  several features will be used to classify the quality of wine, many of them are chemical, so we need to have a basic understanding of such chemicals.

* **fixed acidity**
* **volatile acidity :** Volatile acidity is the gaseous acids present in wine.
* **citric acid :** It is weak organic acid, found in citrus fruits naturally.
* **residual sugar :** Amount of sugar left after fermentation*.*
* **chlorides :** Amount of salt present in wine*.*
* **free sulphur dioxide :** It is used for prevention of wine by oxidation and microbial spoilage.
* **total sulphur dioxide**
* **density**
* **pH :** In wine pH is used for checking acidity
* **sulphates :** Added sulphites preserve freshness and protect wine from oxidation and bacteria.
* **Alcohol**
* **Quality (**Output variable)

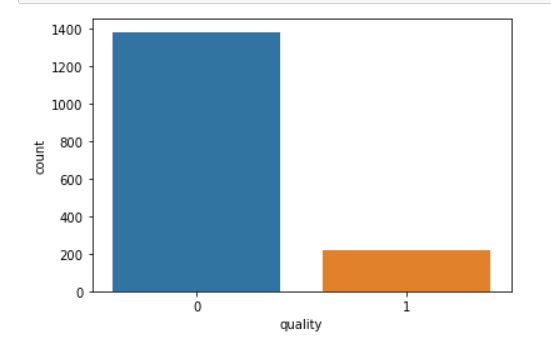
Getting Started:

First and foremost I started with the usuals by importing the necessary libraries like numpy , pandas, matplotlib and seaborn . After that I called dataset using pandas .read \_csv.



I found that When I check the target column(quality) values in between 0 and 10 . so  an interesting thing to do, is to set an arbitrary cut off for the dependent variable (wine quality) at e.g. 7 or higher getting classified as 'good/1' and the remainder as 'not good/0'. For that I find out the total values of the column and made a visualisation of column.

I changed the target column as binary classification. For that create a function and split the column as good/bad. If the quality above seven consider as 0 and else 1.

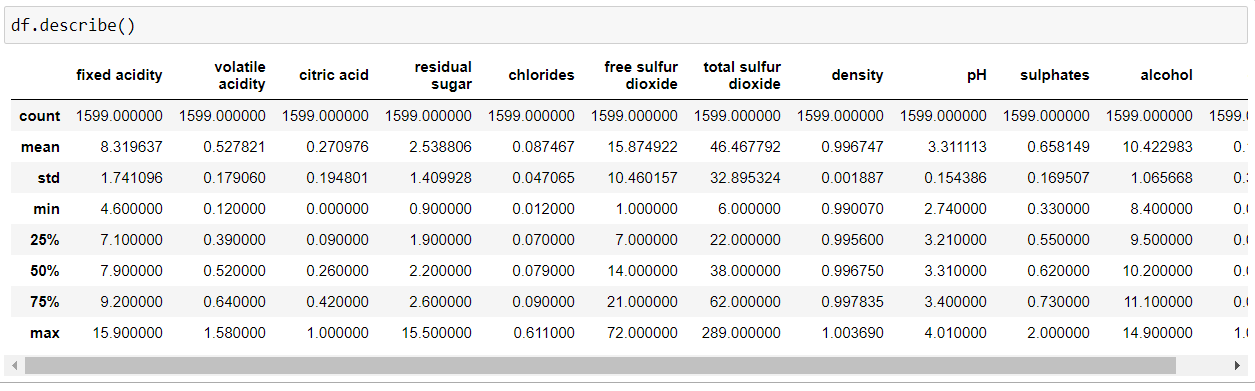


After I found that data imbalance in the target column. Value of zero is 1382 and one is 217. Solved issue of data imbalance at the end .

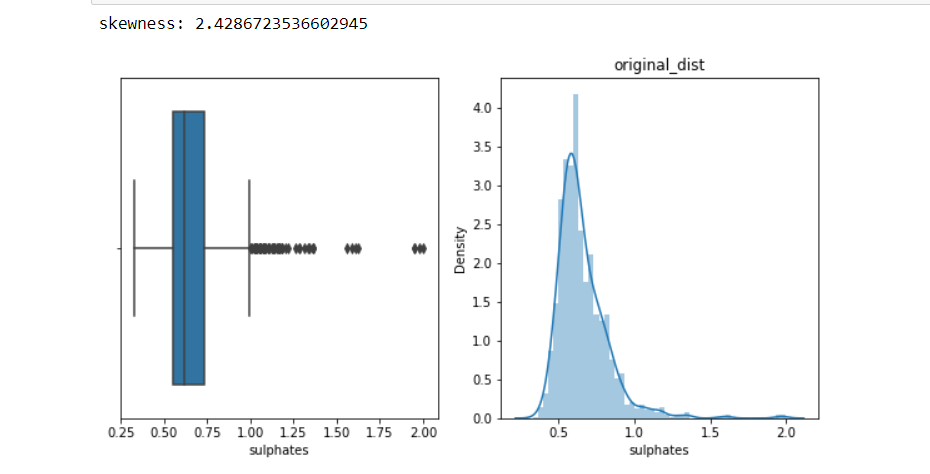
The next step I checked the nan value whether it is or not. There is no missing value present on the dataset.

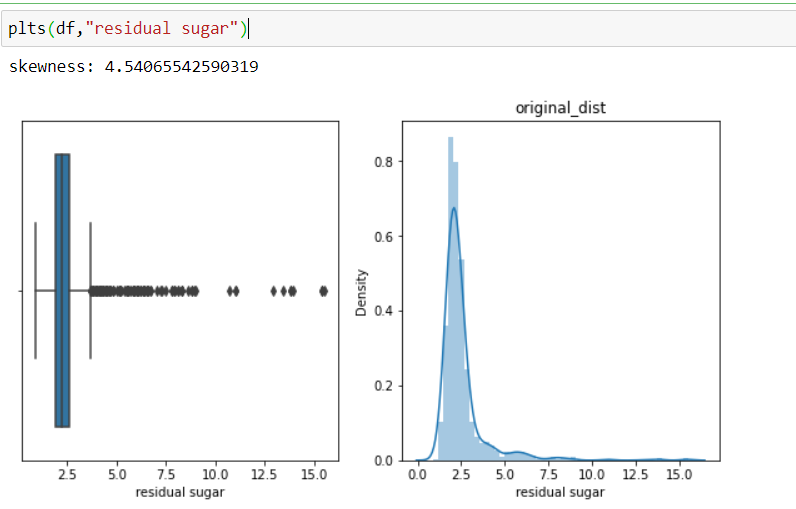
All the columns are float type except target column.

For the better analysis of the data I try to analyse using describe function. Through the table I can understand the outlier of the columns, skewness and spreading of data .



After analysing the data I started cleaning the columns outlier . I used visualization for explaining the data. In other words, we can say that it is a graphic representation of data that is used to find useful information. For understanding the outlier I plot boxplot and for distribution plot distplot.

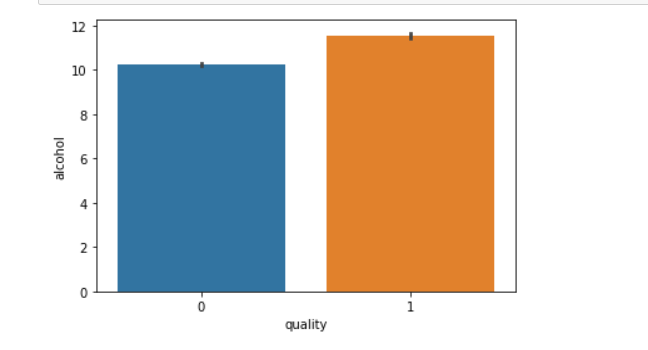




After plotting the graph I can understand that almost all the columns have outlier some columns have high number of outliers like sulphates, pH, density, total sulfur dioxide, chlorides, residual sugar and fixed acidity.

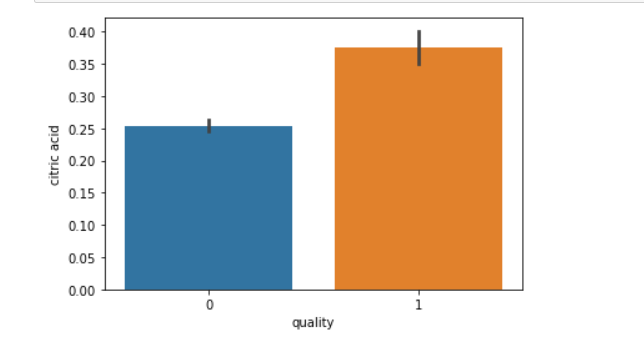
Using IQR I try to find out the outlier and percentile of outlier. Through this I notice that when I remove the outlier huge amount of data might be lost. So did not try to remove outlier.

Count plot:



Through the above graph I can understand that changing the quantity of alcohol will effect the quality of wine. Quantity of alcohol increase from 10 to 12 ,it will also increase good quality of wine.

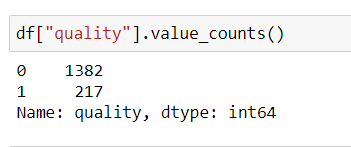
Barplot:



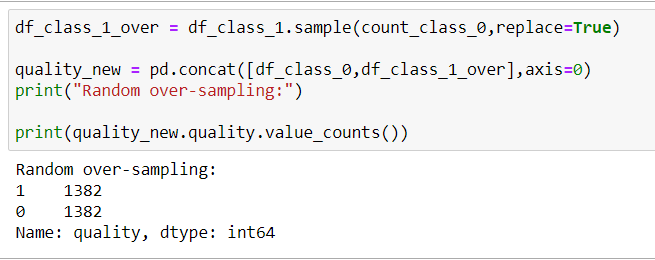
When I plot barplot of citric acid and quality , I learn that increase the level citric acid will also improve the quality of wine.

Handle data imbalance:

For resolving the issue of data imbalance of target column , I used over sampling method. I used this method because this is not much big data set and the different of majority and minority category are also not much high. I already show the graph of target column , through that you also can understand the level of classes.



After applying the random over sample I got balanced data in my label column

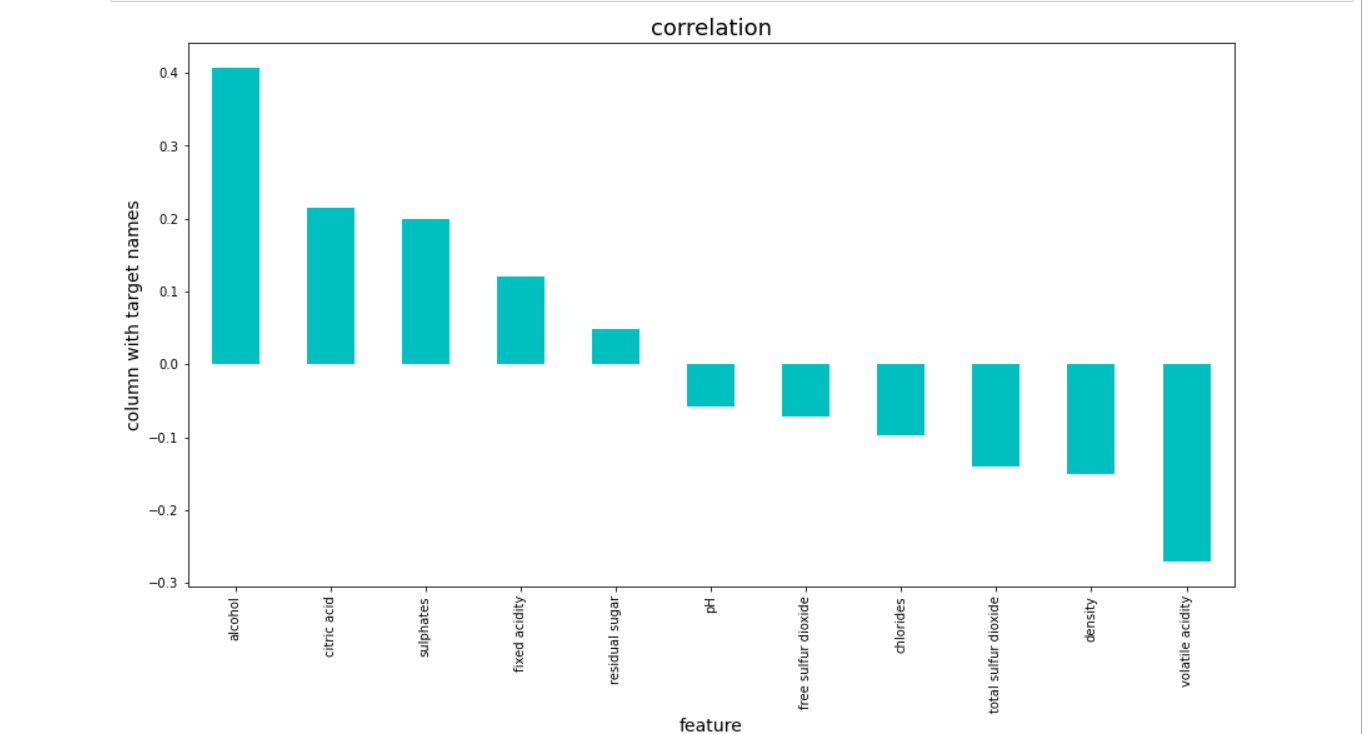


Handling skewness:

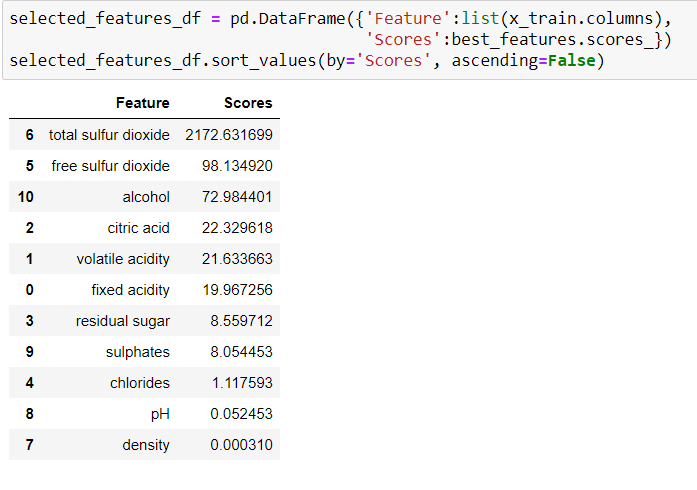
When I check the skewness of data I can see that all the features are skewed except citric acid, density and pH column. I used log transform for handling skewness .

Correlation:

Using the correlation map of independent and dependent features we can understand which feature are more correlated and which are not. If the feature is zero correlated or nearest zero , we can drop the column. And any feature which has equally correlated with each other , we can reduce the feature for better prediction of model. Here alcohol is highly correlated with target and volatile acidity is highly negative correlated with label.



Feature selection also done in the problem. For selecting features I applied univariate selection method. Statistical tests can be used to select those features that have the strongest relationship with the output variable. The scikit-learn library provides the Select K Best class that can be used with a suite of different statistical tests to select a specific number of features. After applying select K Best I got six features. In the below table I took top 6 features.

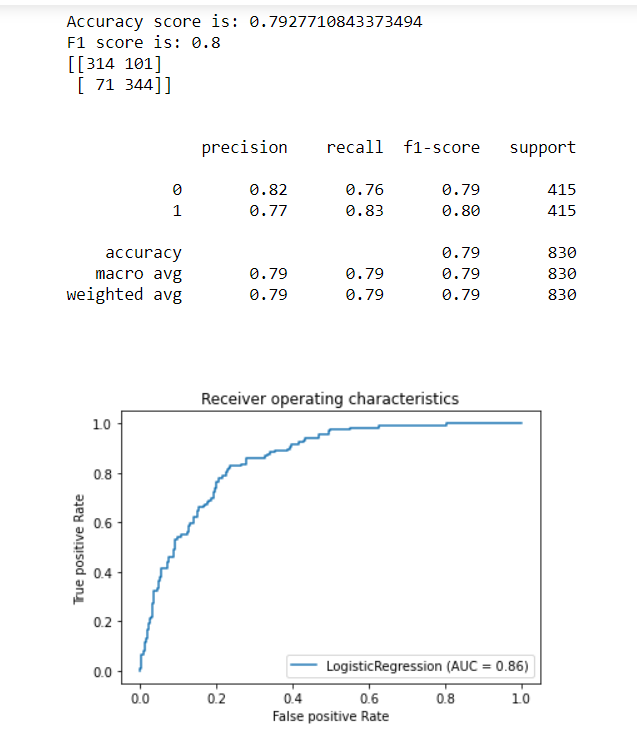


Splitting the data as input and output. Here quality is our output and other feature are our input. After splitting the data as x and y ,I used scaling the x data for single format. For that I applied standerdscaler .

My data are cleaned and ready for the model building. For building the model I used Logistic Regression, Random Forest Classifier and Decision Tree .

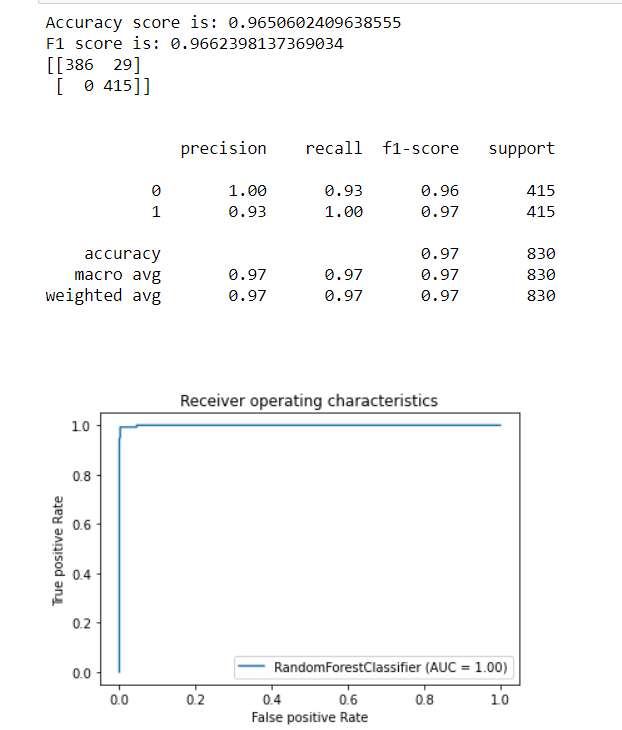
This is the last step where we apply any suitable model which will give more accuracy, compare the above three model’s accuracy and find out that which one is our best model. Details of model :

**Logistic Regression:**



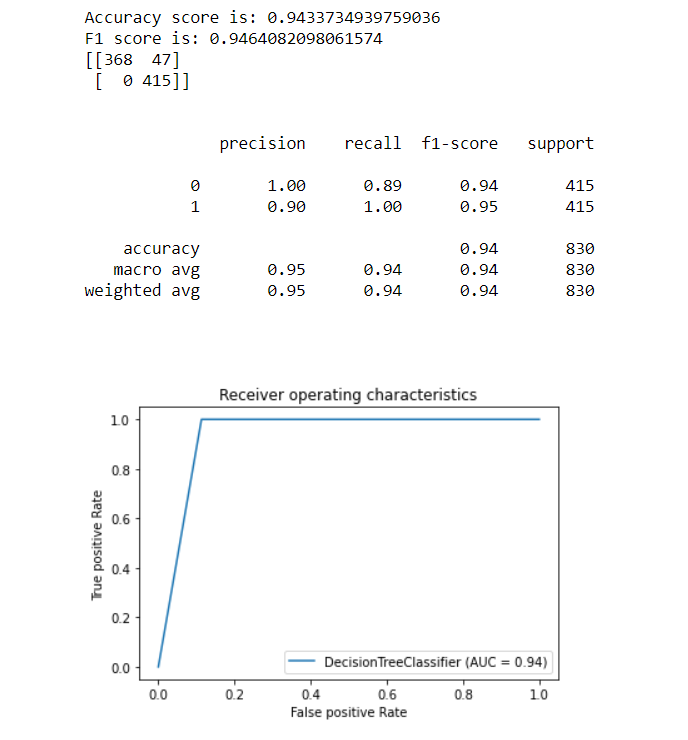
Logistic Regression model’s accuracy is 79 . it’s f1 score and recall value are near by value.

**Random Forest:**



It also perform well. Visualize the model performance using auc and roc curve.

**Decision Tree:**



After getting the accuracy of model I used to chech cross validation for finding the over fitting and underfitting of model.

So finalise the model I find out cross validation score and I finalise the model without over fitting and underfitting.

Compare between accuracy score and cross validation score I can find out my best model.

**Conclusion:**

Logistic Regression is my best model. I can find out through least different of accuracy and cross validation score. This model can predict pretty much better the wine quality when they getting the features above I used. This is one of the best Machin learning problem I have done.

I determined four of the features as the most influential features that is volatile acidity, citric acid, sulphates and alcohol. Good quality wine have lower volatile acidity, higher alcohol and high value of citric acid. For improve the accuracy of the model need more closely related features like wine type, year of harvest etc.

For improving the performance of the model I tried to model selection algorithm as Gride search CV . After tried tuning I got better accuracy.

Finally I saved my model in job lib

