Red Wine Quality Prediction Using Classification Techniques

In this step-by-step process of Predicting the quality of red wine, you will see the importance of statistics in transforming, interpreting, analyzing, and visualizing data.

Problem Statement

The dataset is related to red and white variants of the Portuguese "Vinho Verde" 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.

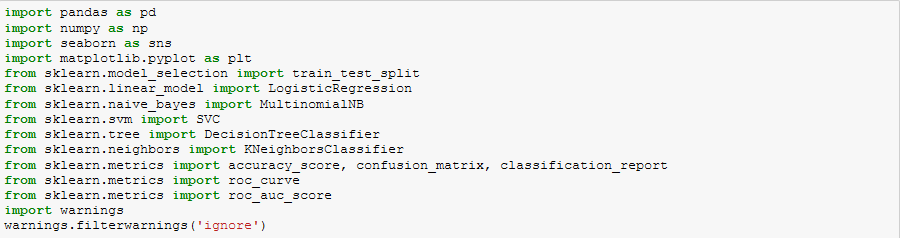
Now, a brief overview of the Red Wine Quality Dataset.

The main aim of the red wine quality dataset is to predict which of the physiochemical features make good wine. With 11 variables and 1 output variable (quality) given, let us examine the role of each of these features:

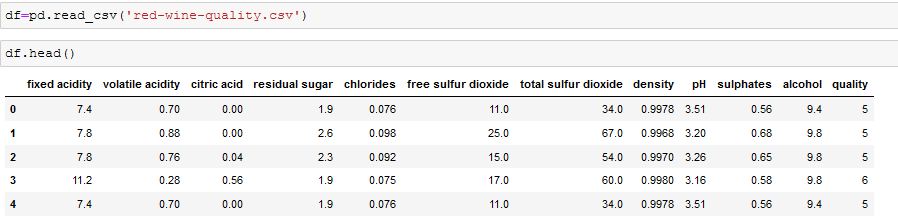
1. ***Fixed Acidity***: are non-volatile acids that do not evaporate readily
2. ***Volatile Acidity***: are high acetic acid in wine which leads to an unpleasant vinegar taste
3. ***Citric Acid***: acts as a preservative to increase acidity. When in small quantities, adds freshness and flavor to wines
4. ***Residual Sugar:*** is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between sweetness and sourness. It is important to note that wines > 45g/ltrs are sweet
5. ***Chlorides***: the amount of salt in the wine
6. ***Free Sulfur Dioxide:*** it prevents microbial growth and the oxidation of wine
7. ***Total Sulfur Dioxide***: is the amount of free + bound forms of SO2
8. ***Density***: sweeter wines have a higher density
9. ***pH***: describes the level of acidity on a scale of 0–14. Most wines are always between 3–4 on the pH scale
10. ***Alcohol***: available in small quantities in wines makes the drinkers sociable
11. ***Sulphates***: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
12. ***Quality***: which is the output variable/predictor

Now we have a basic knowledge of various factors that influence the quality of good wine.

Importing the Dataset



Getting the Data



Data Cleaning And Analysis

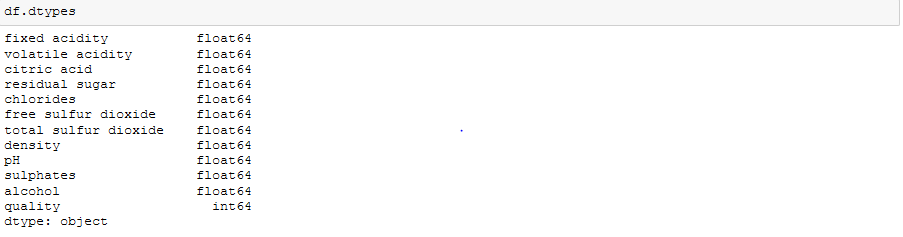
Let’s explore some features in the dataset

*1. Size of the dataset:*



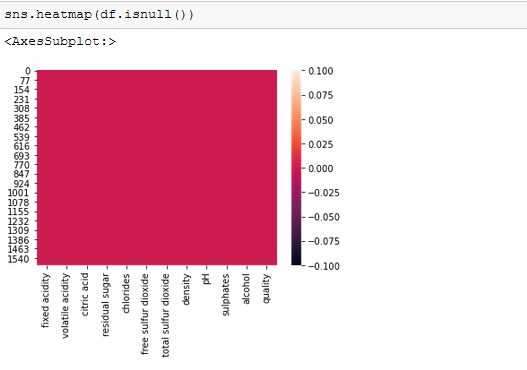
There are 1599 rows and 12 columns in our dataset.

*2. Type of the Data:*



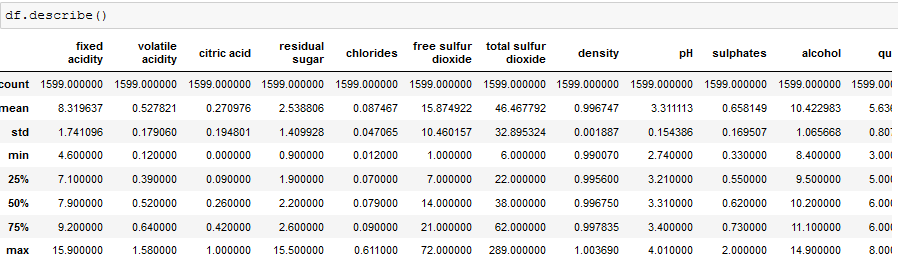
All the features are numeric. There is no categorical feature in our dataset.

*3. Checking for Null Values:*



We can see there are no null values in our dataset.

*4. Describe the Dataset:*

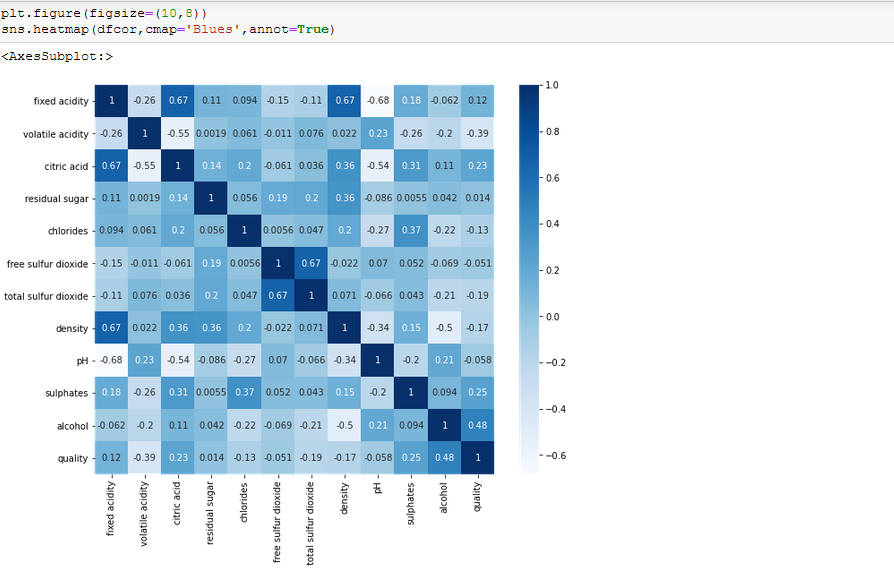


Difference in mean and median of columns fixed acidity, free sulphurdioxide,total sulphur dioxide,sulphates suggests data is skewed in these columns.

Difference between 75% quartile and maximum value reveals there are outliers present in the columns residual sugar,free sulphuedioxide & total sulphurdioxide.

Standard deviation is high in free sulphurdioxide,total sulphurdioxide also suggests data is widely spreaded.

*5. Correlation with target variable*



*P{XY} = corr(X,Y)*

* *1.0 ≤P (XY) ≤+1.0*

Where; 0: No relationship

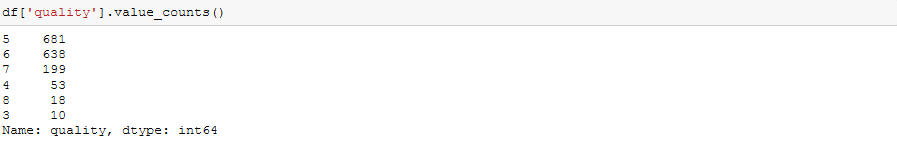
(+/-)0.6: Moderate relationship

(+/-)0.8: Fairly Strong relationship

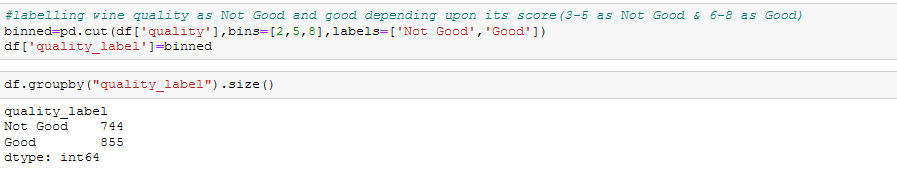
(+/-)1: Perfect Relationship

from the above map, we can conclude that pH, free sulfur dioxide and residual sugar have least correlation with the quality.

*6. Exploring Target Variable:*



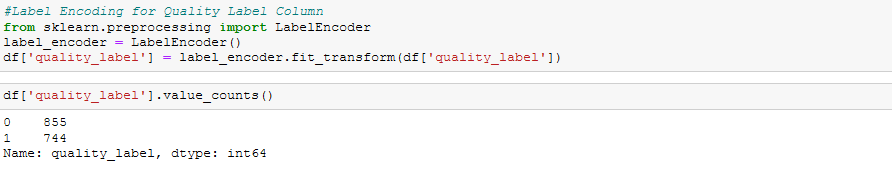
Value of quality ranges from 3 to 8, data is quite imbalanced. We will try to make a precise category for better evaluation.



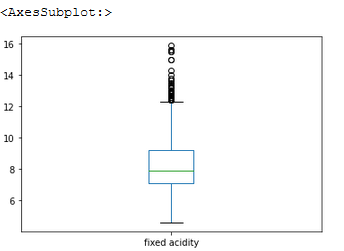
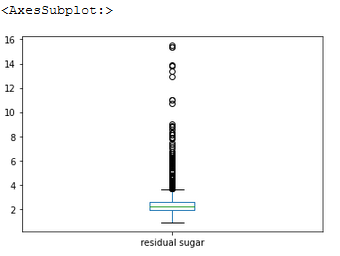
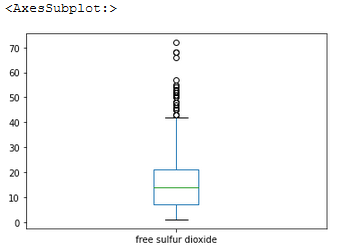
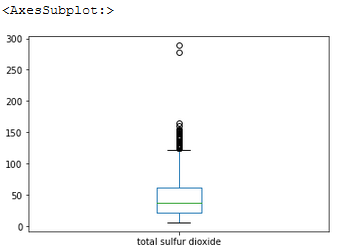
Now we can see our target class is properly balanced.

*7. Label Encoding the target variable*

We will perform label encoding for quality\_label column using label encoder. We will import LabelEncoder from sklearn.preprocessing.



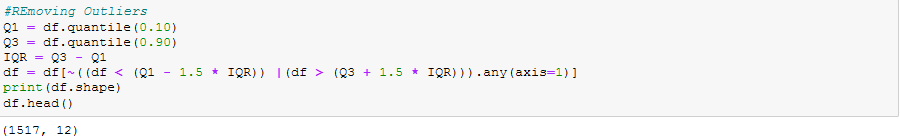
*8. Outlier Detection Using Box-Plot*

*  *

We can see outliers are present in all the four features of our dataset. We will try to remove them in further analysis.

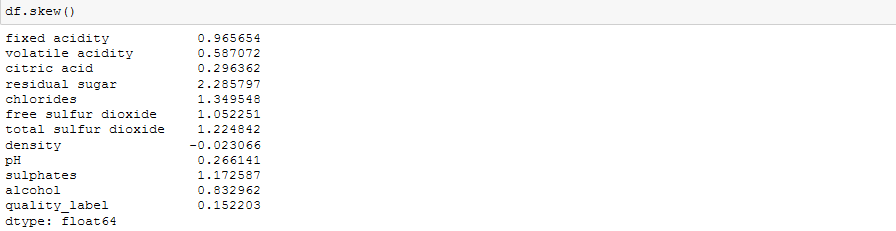
*9. Outlier Removal using IQR Method*

The technique is by using the rule of IQR (Interquartile Range) where everything below Q1 − 1.5 \* IQR and above Q3 + 1.5 \* IQR is considered as outliers. The advantage is that it does not depend on the variable’s distribution.

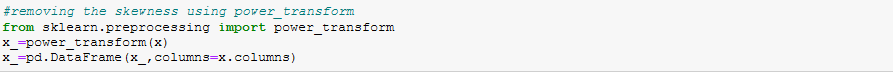


After removing outliers, we have 1517 rows and 12 columns in our dataset.

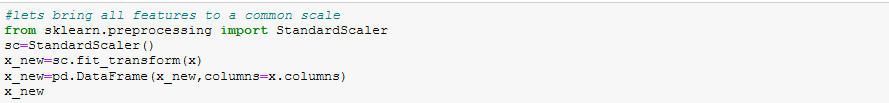
*10. Checking the Skewness*



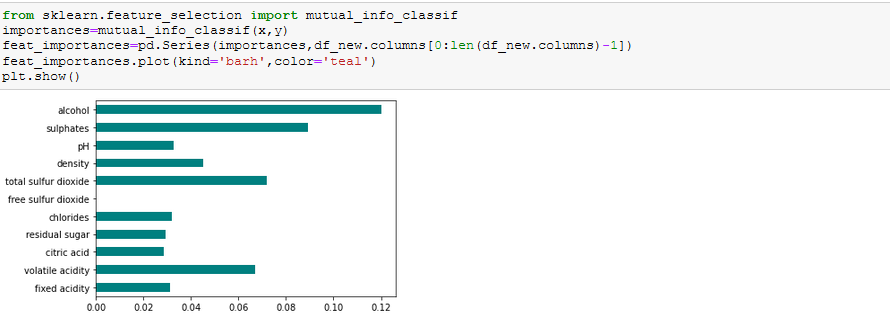
We can see residual sugar, chlorides, free sulfurdioxide, total sulphurdioxide and sulphaes are highly skewed. So, we will remove the skewness using power transform.



*11. Scaling the data*



Feature Selection



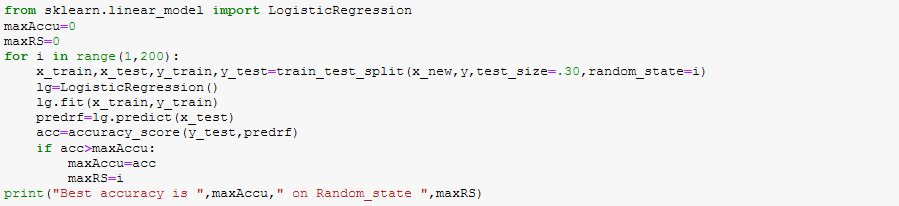
Visualizing the feature importance we can see the top features are alcohol, volatile acidity, sulphates, density, totalsulfurdioxide, and they will improve the wine quality.

We will drop the features which are least important in improving the wine quality.

Building Machine Learning Models:

Now we will train several Machine Learning models and compare their results. Note that because the dataset does not provide labels for their testing-set, we need to use the predictions on the training set to compare the algorithms with each other. Later on, we will use cross validation.

Finding the Best Random State

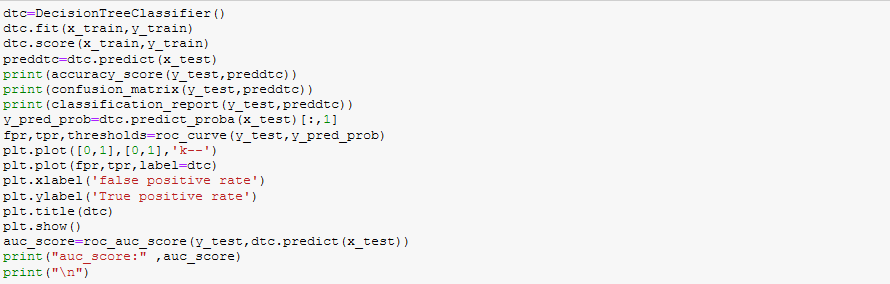


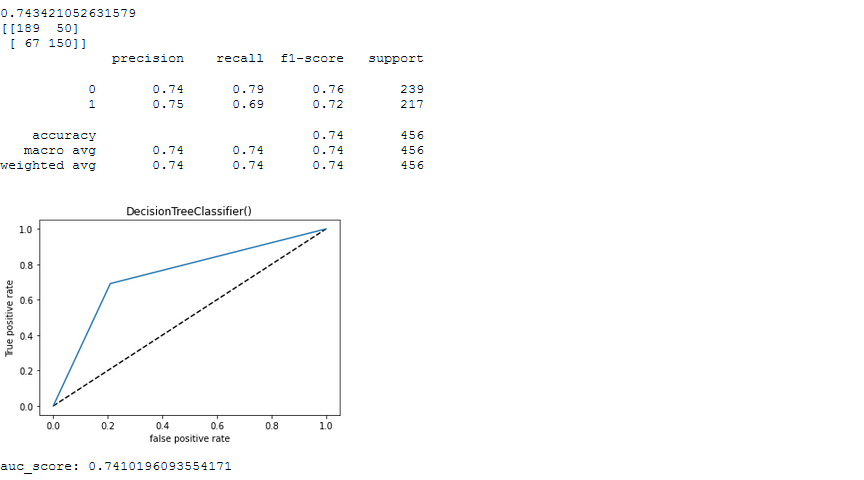
Best accuracy is 0.7982456140350878 on Random\_state 190

**Train and Test Data Split**

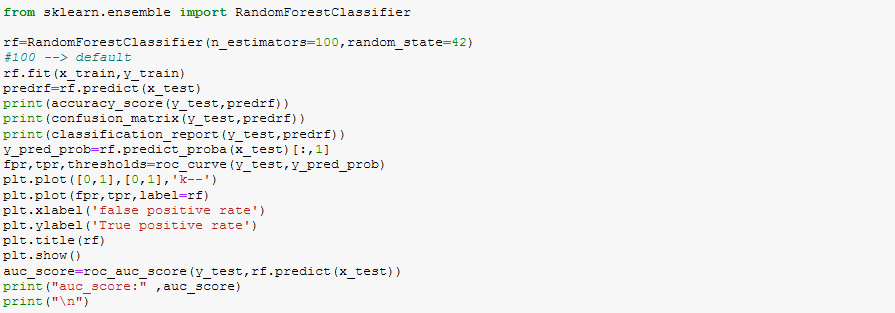
****

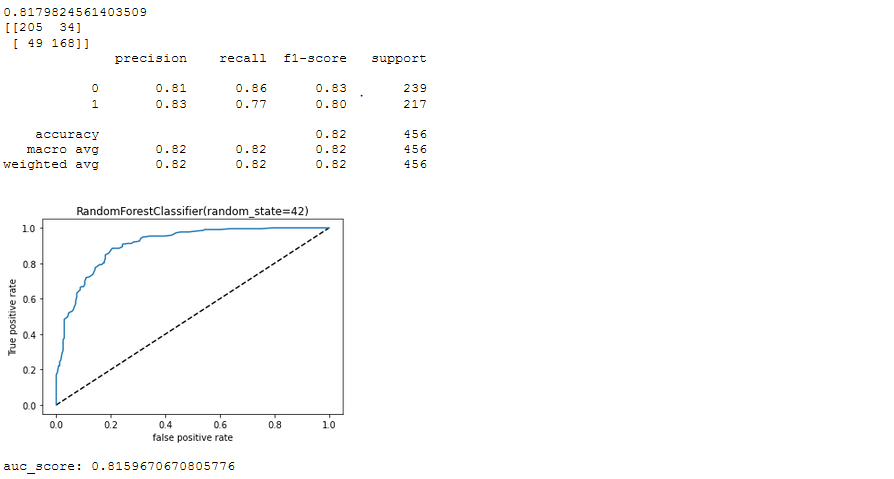
Decision Tree



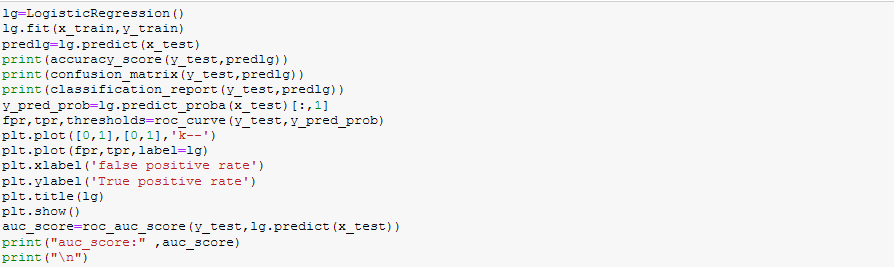


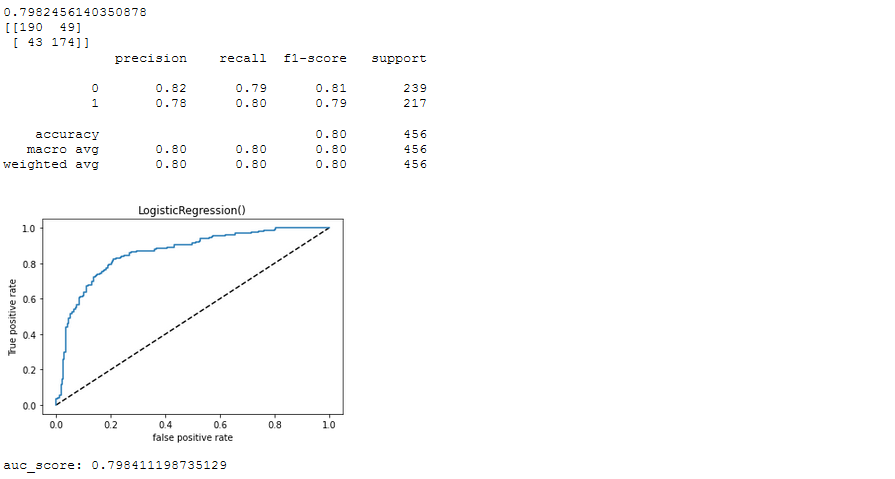
Random Forest



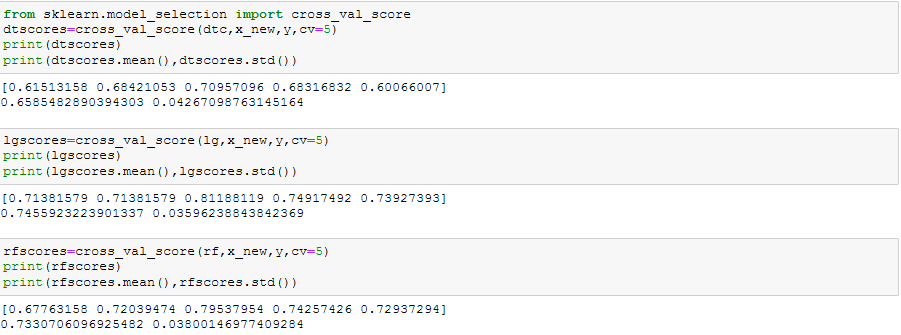


Logistic Regression





Performing Cross-Validation:



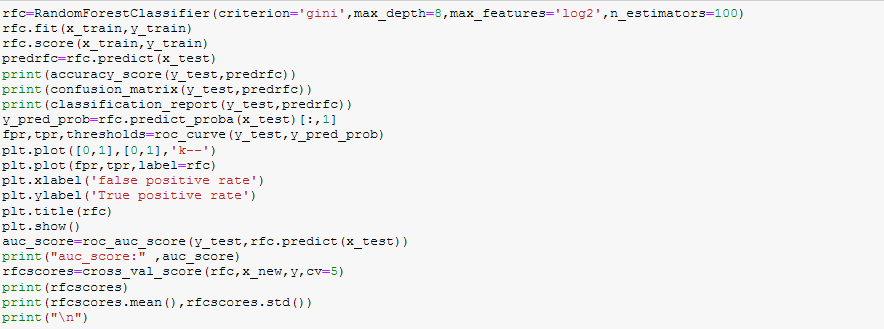
# *Which is the best Model ?*

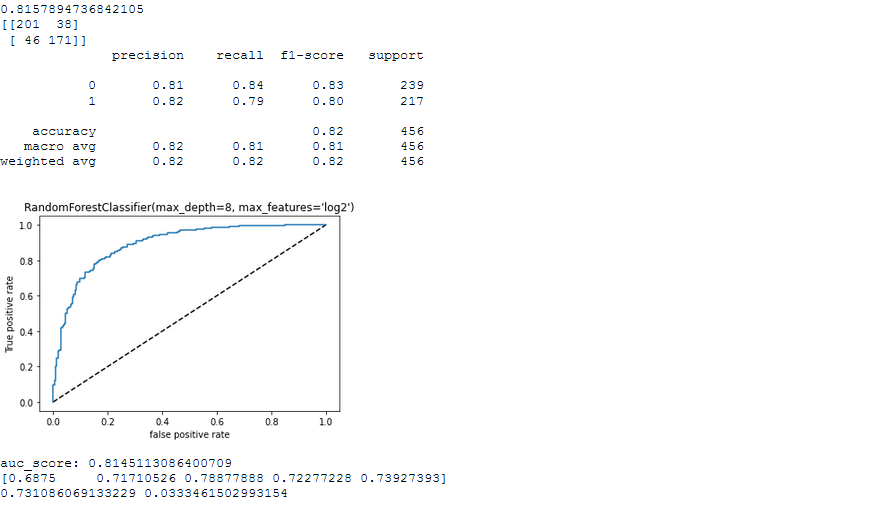
As, Logistic Regression and Random Forest Classifier have the least difference between accuracy and cross validation score, we will perform hyper parameter tuning on them to decide which is best model.

Hyper-Parameter Tuning

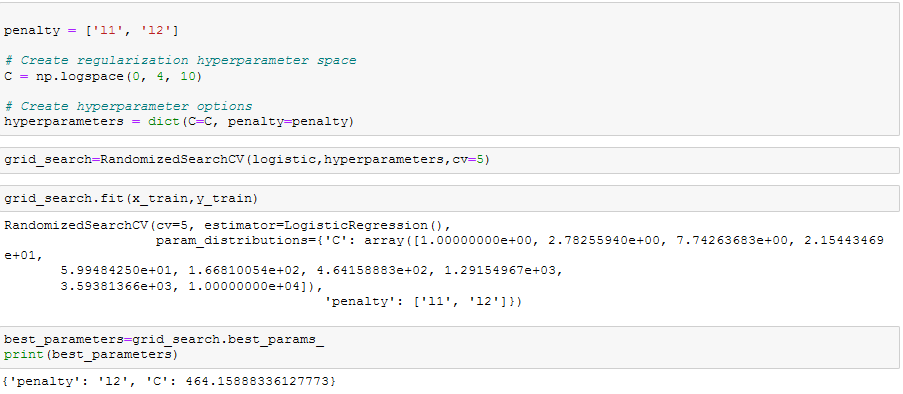
*Random Forest Hyper-Parameter Tuning*

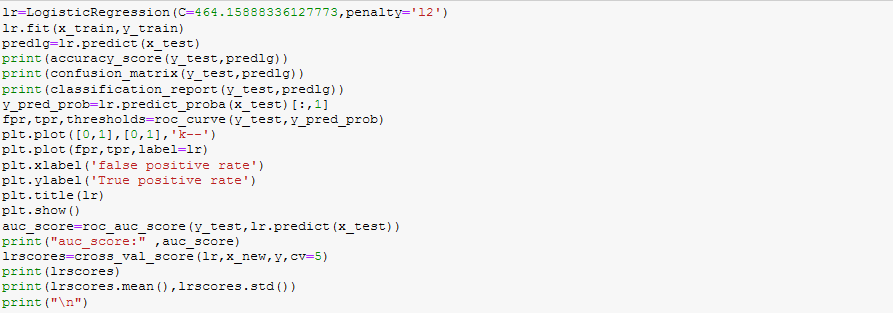


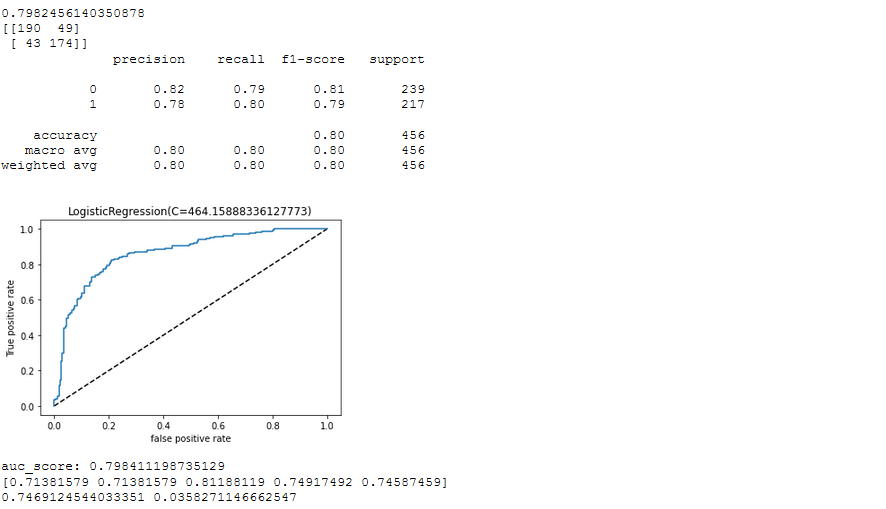




*Logistic Regression Hyper-Parameter Tuning*







Selecting Our Best Model

After hyper parameter tuning, Random Forest Classifier have the best AUC score and cross validation score thus, we will select it as our best model.

## *What is Random Forest ?*

Random Forest is a supervised learning algorithm. Like you can already see from it’s name, it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. With a few exceptions a random-forest classifier has all the hyper parameters of a decision-tree classifier and also all the hyperparameters of a bagging classifier, to control the ensemble itself.

The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model. Therefore when you are growing a tree in random forest, only a random subset of the features is considered for splitting a node. You can even make trees more random, by using random thresholds on top of it, for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

# Summary

We started with the data exploration where we got a idea for the dataset, inspected about missing fields and learned which features are important using correlation matrix. During this process we used Seaborn and Matplotlib library to do the visualizations. During the data preprocessing, we computed outliers and removed skewness in data, converted categorical features into numeric ones. We checked which features are highly important for quality of the wine. We scaled the data for un-biased model evaluation. After that, we trained 3 different machine learning models, picked two of them (Random Forest & Logistic Regression) and applied hyper parameter tuning on it. Then we discussed how random forest works, took a look at the importance it assigns to the different features and tuned its performance by optimizing its hyper parameter values. Lastly, we looked at its confusion matrix and computed the model’s precision, recall, f1-score . We also computed ROC-AUC curve and AUC Score.