1) Problem Definition:-

In this blog we will check the quality of Wine. Whether it’s good or bad.

For that study we have dataset related to red and white variants of the Portuguese "Vinho Verde" wine.  This dataset has the fundamental features which are responsible for affecting the quality of the wine. By the use of several Machine learning models, we will predict the quality of the wine. In this dataset, classes are ordered, but it was not balanced. Here, red wine instances are present at a high rate and white wine instances are less than red.

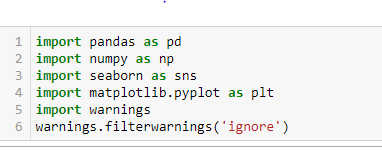
With 11 variables and 1 output variable (quality) given, let us examine the role of each of these features:

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. **Sulphate**: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
12. **Quality:** which is the output variable/predictor

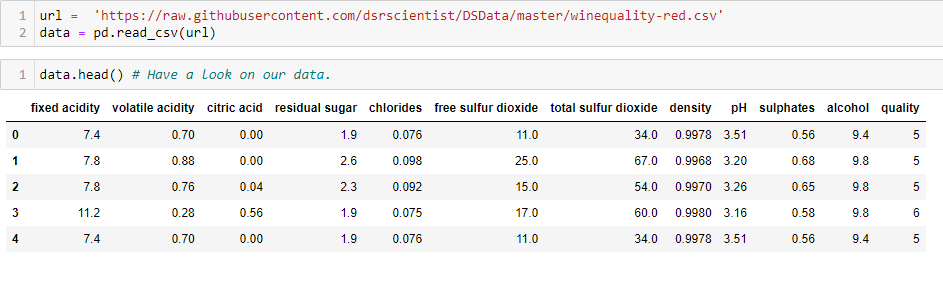
Now we have a basic knowledge of various factors that influence the quality of good wine, let’s go for Data Analysis.

2) Data Analysis:-

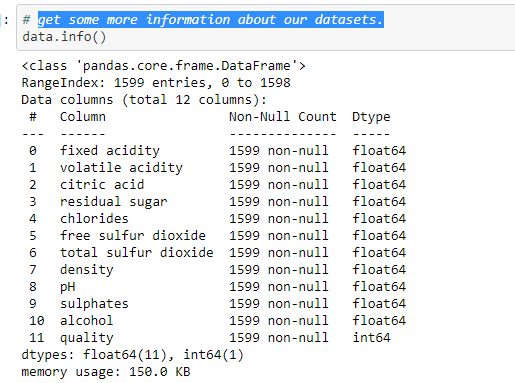
**Let’s start with importing the required modules**



**Importing dataset and see what is inside the data set by seeing the first five values of dataset by head () command.**

 There are two things, features and labels. Features are the part of a dataset which are used to predict the label. And labels on the other hand are mapped to features. So, if we analyses this dataset, since we have to predict the wine quality, the attribute quality will become our label and the rest of the attributes will become the features.

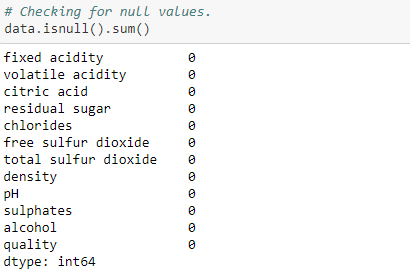
**Checking size of our dataset using .shape command.**



The shape of the data (1599, 12) which shows there are 1599

Rows and 12 columns in the data.

**Checking for null values.**

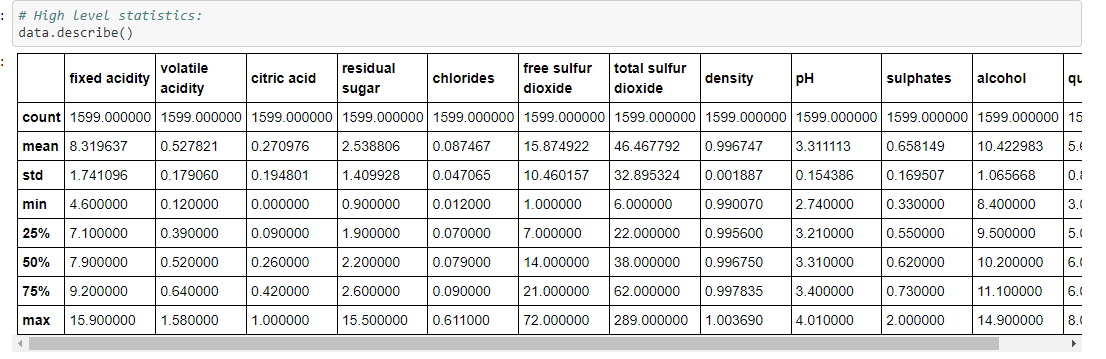


WE can see there is no null values in our dataset.

Getting some basic information about our dataset like data types of each column and memory usage of the entire data.

**Data Description and Exploratory Visualizations**

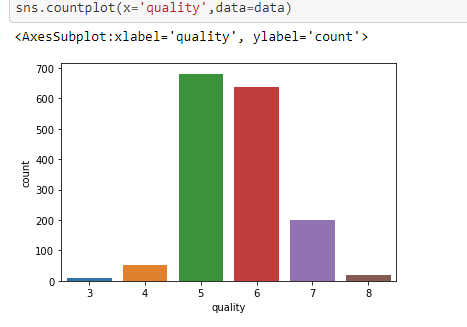
Let’s check data description using .describe ()



As we can see here, mean value is less than the median value of each column.

There is a large difference between the 75th% tile and max values of residual sugar, free sulfur dioxide & total sulfur dioxide.

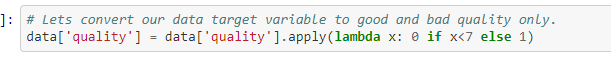
**Let’s plot and check unique values count in quality.**



**Quality** has a high number of values in categories 5, 6 and 7.

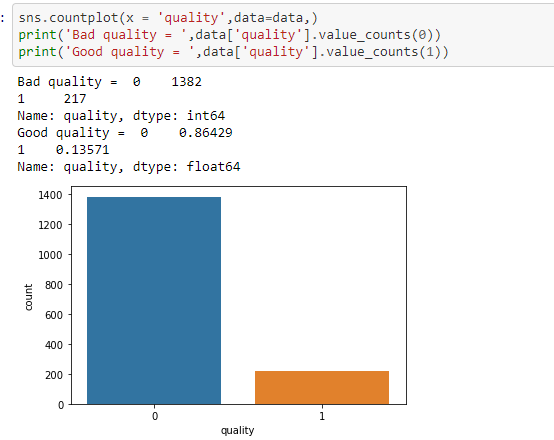
Only a few observations are there for the categories 3 & 9.

We can see our Label contains 6 different categories for wine quality but out predict is only for Bad and Good. So we will apply lambda function to convert our Label into 0 and 1 whereas if the value is less than 7 the quality of wine is bad, and if the quality is above 7 then wine quality is good.



We have applied Lambda function on our Label column.

**Plotting quality counts.**



We can see there is Class **Imbalance** problem in our target variable.

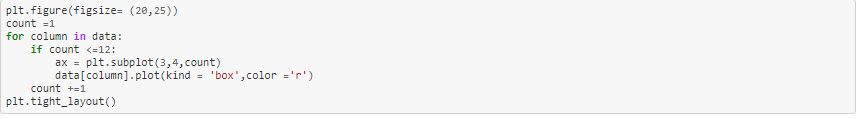
**Try to gather more information regarding data distribution.**

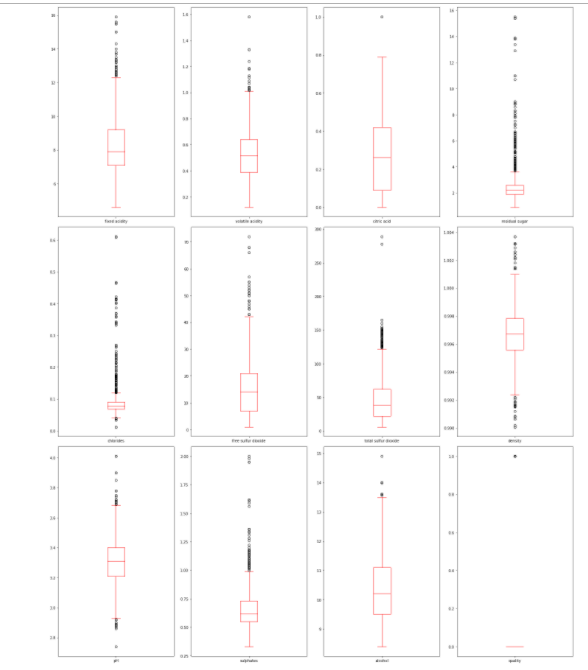


As we can see:-

* In fixed acid data is right skewed.
* Volatile acidity also show some right skewness.
* Citric acid data is right skewed.
* In Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide and Sulphate columns data is highly right skewed.

**As we have skewness in our dataset let’s check for Outlier as well by plotting Box plot**.

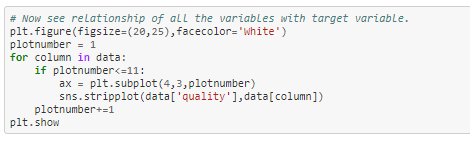


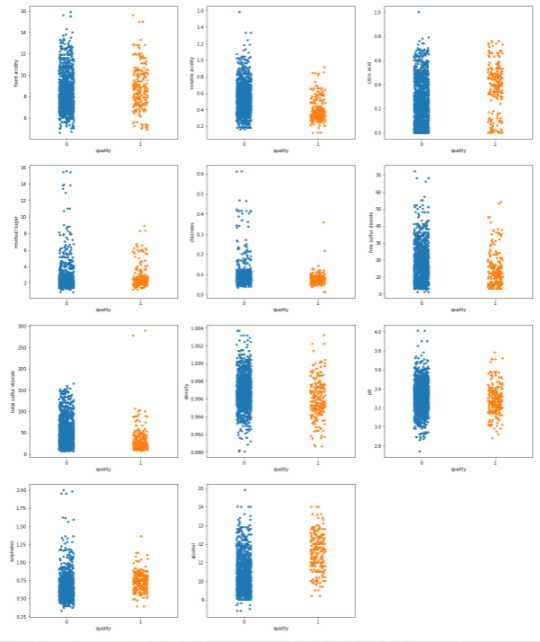


From the above boxplot we can clearly see that our maximum columns have outlier in it.

So we have one more task we have to remove the outliers.

**Let’s plotting relationship between Label and all the features.**

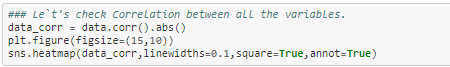


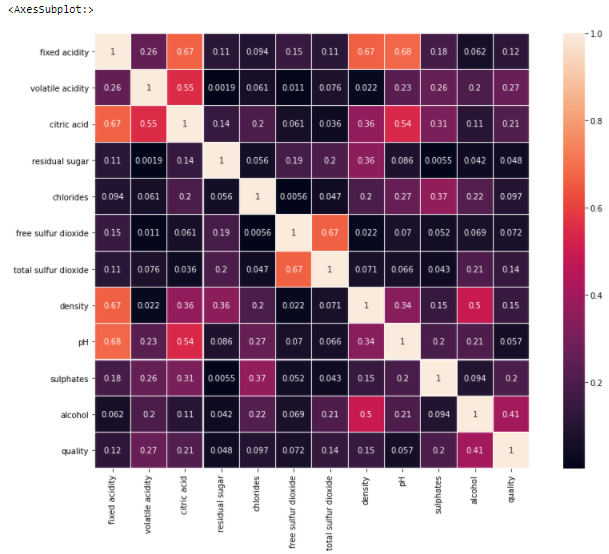


We can see as Sulphates and cholrides amount increasres our win quality decreases.

**Correlation**

Let’s take a look at some of most significant correlations. It is worth remembering that correlation coefficients only measure linear correlations.





Density has a strong positive correlation with residual sugar, whereas it has a strong negative correlation with alcohol.

PH & fixed acidity has negative correlation.

Density & fixed acidity has positive correlation.

Citric acid & fixed acidity has positive correlation.

Citric acid & volatile acidity has negative correlation.

Free sulphur dioxide & total sulphur dioxide has positive correlation.

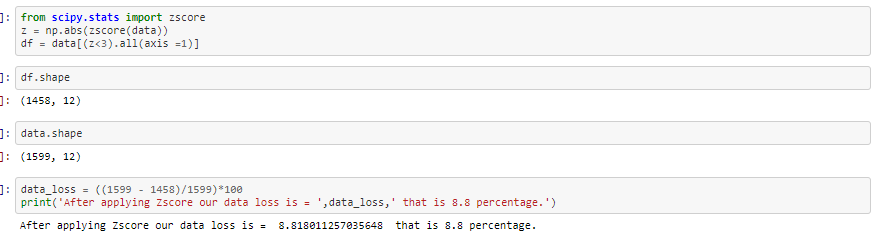
3) EDA Concluding Remarks:-

* The dataset does not feature any missing or erroneous data values, and all features are of the correct data type.
* We can see there is Class Imbalance problem in our target variable.
* There is skewness in our dataset.
* Also we have Outliers in most columns.
* Alcohol is 41% related to wine quality. Also density, PH and citric acid have 67%, 68% and 67% relationship with fixed acidity respectively. Total sulfur dioxide and free sulfur dioxide shows 67% relationship.

4) Pre-processing Pipeline:-

We undertake data pre-processing steps to prepare the datasets for Machine Learning algorithm implementation.

**Applying z-score to remove outliers.**



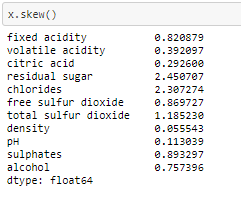
After removing outliers we check for data loss and found only 8.8% data loss.

## **Separating Independent variables and Target variable**



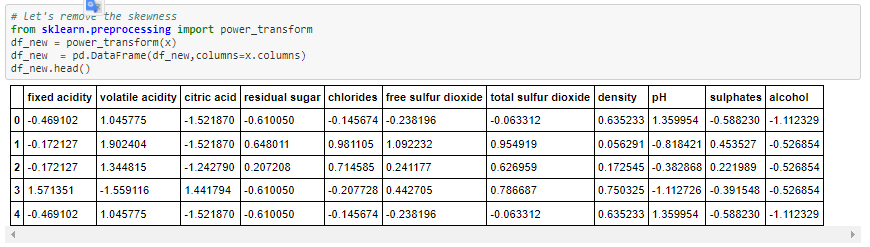
In X we stored Features and in Y we store our label.

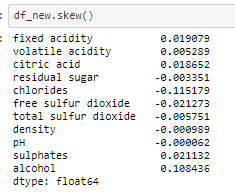
**Checking skewness in our X features.**



As a general rule of thumb: If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric the above output shows the skewness score for all the features.

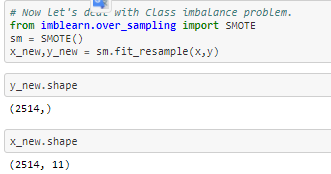
**Let’s apply skewness removing technique to remove skewness from each features.**





As we can see skewness is removed from all the columns.

**Now let’s deal with class imbalance problem.**

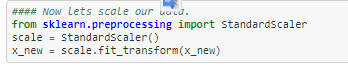


By applying class Imbalancing technique we double the dataset for good quality.

**Standardization**

Standard Scalar follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance

Before implementing models we need to normalize our features to remove Biasness. For that we use Scaling technique.

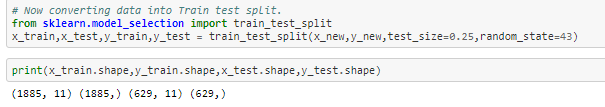


Scaled our X data so that our model will not Bias to any of the variable.

**Train - Test Split**

Prior to implementation or applying any Machine Learning algorithms, we must decouple training and testing data from our master dataset

Applying train test split on x and y.

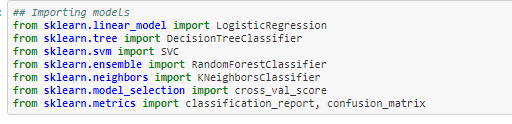
We have divided data into training and testing for model building we will use 75% data for training and use 25% for testing purpose.

5) Building Machine Learning Models:-

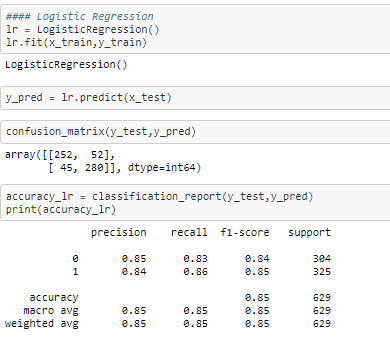
As we know this is classification problem we are importing 5 classification models to train and test our dataset. Also importing model selection and Metrics to evaluate our model accuracy.

**Classification Accuracy** is the number of correct predictions made as a ratio of all predictions made. It is the most common evaluation metric for classification problems.

A **confusion matrix** is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

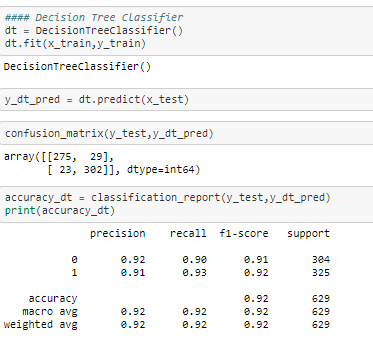


**Logistic Regression**



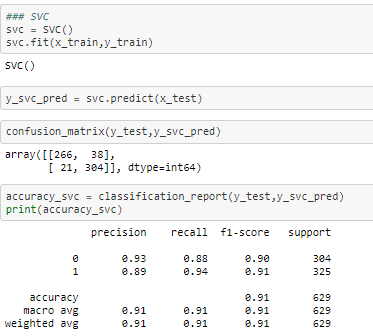
We can see accuracy for Logistic regression model is 85%

**Decision Tree Classifier**



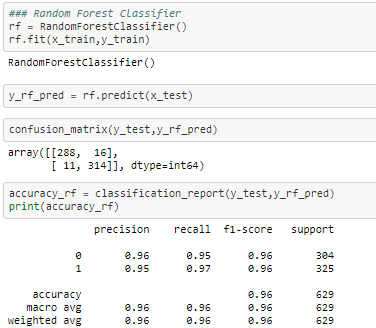
The accuracy for Decision tree classifier is 92%

**SVC**



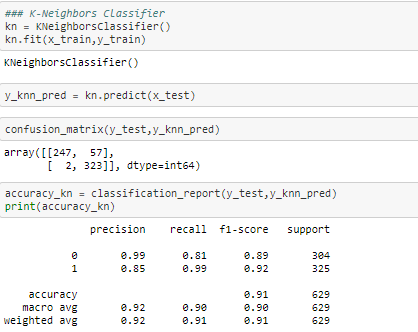
Accuracy for SVC is 91%

**Random Forest Classifier**



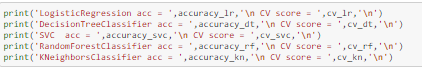
The accuracy for Random Forest classifier is 96%

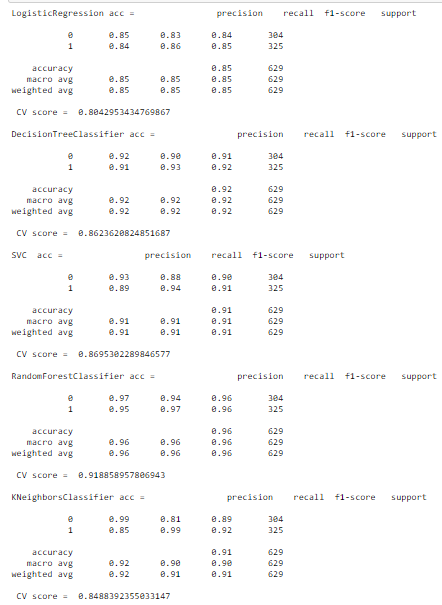
**KNN-Neighbors**



The accuracy for KNN-Neighborsis 91%

**Cross-validation** is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. Let’s check cross validation score for all the models.

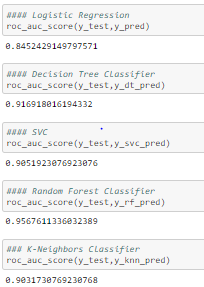




Out of 5 model Random Forest Classifier is our best model, because it’s given as accuracy 96% and Cross validation score 91.92%. Therefore it’s our best model.

**ROC\_AUC\_Scrore and ROC graphs.**

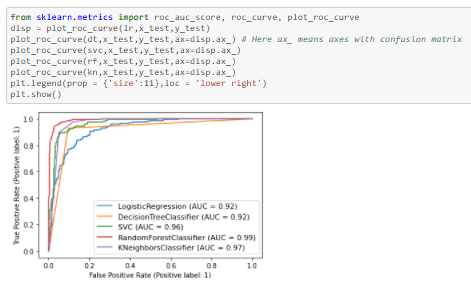
The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



Above are the AUC ROC score which shows higher for Random Forest classifier.

**ROC Graphs**

AUC — ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. The red line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).

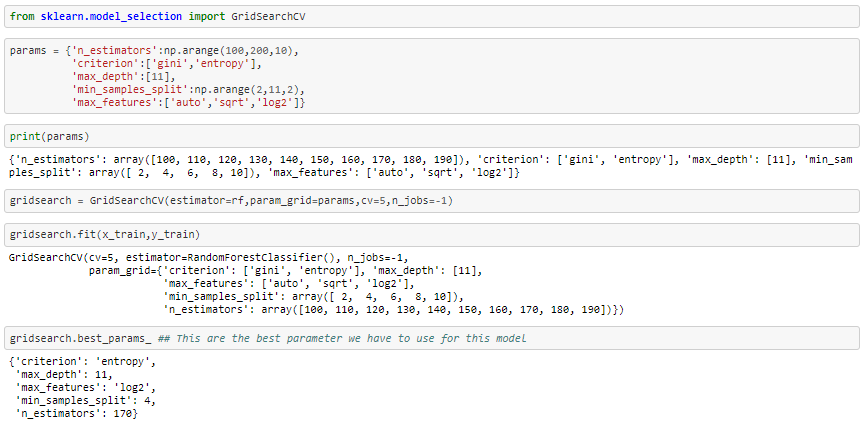


As shown above, the fine-tuned Random forest classifier model showed a higher AUC score compared to other models.

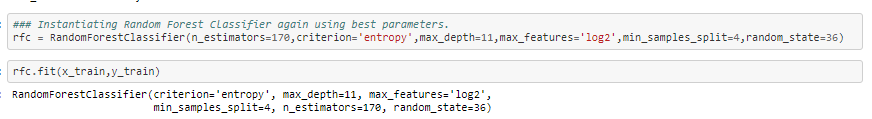
**Hyper parameter tuning using Grid Search CV**

Hyper parameters tuning is crucial as they control the overall behavior of a machine learning model. Every machine learning models will have different hyper parameters that can be set. A hyper parameter is a parameter whose value is set before the learning process begins. Grid search is arguably the most basic hyper parameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyper parameter values provided, evaluating each model, and selecting the architecture which produces the best results**.**

**Let’s import Grid search CV from SK-learn library**

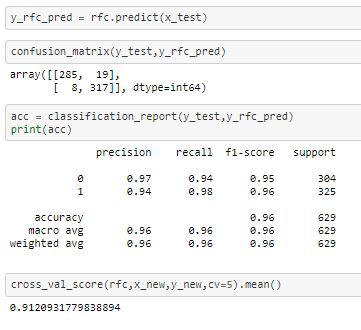


We get the best parameters by Grid search CV. Now let’s instantiating our model again with parameters calculated by Hyper parameter.



We have trained our new model on our training dataset.

Let’s predict our test data using new model and evaluate the model using cross validation and classification matrices.



After parameter tuning on Random forest classifier our accuracy is 96% and our cross validation score is 91.2%.

We didn't see any change in accuracy and cross validation score. Additionally, it is not even guaranteed to find the best solution, often aliasing over the best configuration. After using Grid search CV.

6) Concluding Remarks:-

Based on the all the graphs and plots we come to an conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not have change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection.

For classifying the wine quality, we have implemented multiple algorithms, namely:

1. Logistic Regression
2. Decision Tree Classifier
3. Support Vector Classifier
4. Random Forest
5. K-Neighbors Classifier

We were able to achieve maximum accuracy using random forest of 96%. Logistic Regression giving an accuracy of 85%. Decision Tree Classifier has an accuracy of 92%. SVC giving accuracy of 91%. And K-Neighbors Classifier as 91%.