***Prediction of Red Wine Quality***

**Introduction:**

## “ Wine is the most healthful and most hygienic of beverages “

*– Louis Pasteur*

Yes, if you think deep down then you just notice that we are discussing wine, above quote seems to be right because all over the world wine was soo popular among people, and 5% of the population doesn’t know what is wine? sounds good.

We definitely came across the fruit **graphs**, which is soo sweet on the test but graphs are not just to eat, they are used to make different types of things. Wine is one of them **Wine is an alcoholic drink that is made up of fermented grapes**. If you have come across wine then you will notice that wine has also their type they are red and white wine this was because of different varieties of graphs.

You are shocked to hear that the worldwide distribution of wine is 31 million tonnes which were huge in number.

### What if you think about the quality of wine, how can you differentiate the wine according to their quality? The big question arises.

According to experts, the wine is differentiated according to its **smell**, **flavor**, and **color,**but we are not a wine expert to say that wine is good or bad. What will we do then? Here’s the use of **Machine Learning** comes, yes you are thinking to write we are using machine learning to check wine quality. ML have some techniques that will discuss below:

Now, we start our journey towards the prediction of wine quality, as you can see in the data that there is red and white wine, and some other features. Let’s start :

**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.

**Description of Dataset:**

If you download the dataset, you 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.

* ***volatile acidity :****Volatile acidity is the gaseous acids present in wine.*
* ***fixed acidity :****Primary****fixed acids****found in wine are****tartaric****,****succinic****,****citric****, and****malic***
* ***residual sugar :****Amount of sugar left after fermentation.*
* ***citric acid :****It is weak organic acid, found in citrus fruits naturally.*
* ***chlorides :****Amount of salt present in wine.*
* ***free sulfur dioxide :****So2 is used for prevention of wine by oxidation and microbial spoilage.*
* ***total sulfur dioxide***
* ***pH :****In wine pH is used for checking acidity*
* ***density***
* ***sulphates****:    Added sulfites preserve freshness and protect****wine****from oxidation, and bacteria.*
* ***alcohol :****Percent of alcohol present in wine.*

Rather than chemical features, you can see that there is one feature named **Type**it contains the types of wine we here discuss on **red** and **white** wine, the percent of red wine is greater than white.

**TYPE OF PROBLEM:**

The above problem is a clear classification problem as we need to classify whether the quality is 0 to 10. dependent variable (wine quality) at e.g. 7 or higher getting classified as 'good/1' and the remainder as 'not good/0'.

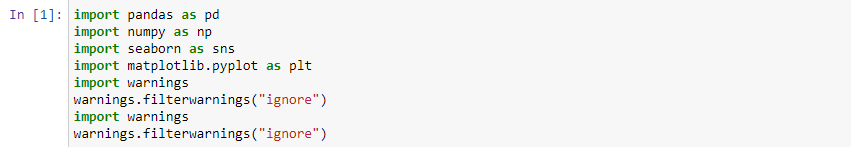
So this can be solved by any of the classification techniques like

1. Logistic Regression .
2. Decision Tree Algorithm.
3. Random Forest Technique.

I have mentioned only few. We will be dealing with each of techniques later in this blog.

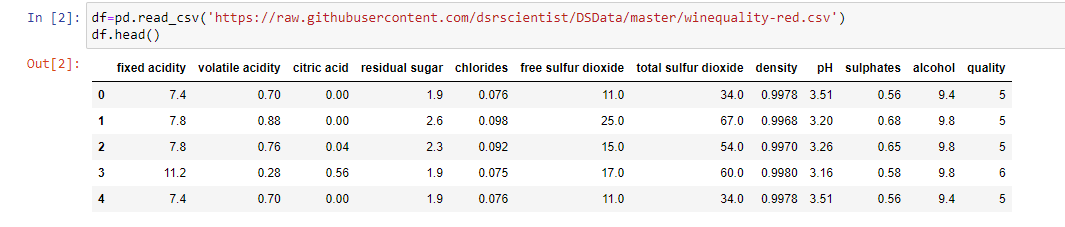
### ***Importing modules:***

Now let me walk through the code. Firstly I just imported the necessary packages like pandas, numpy, seaborn etc. so that i can carry the necessary operations further.



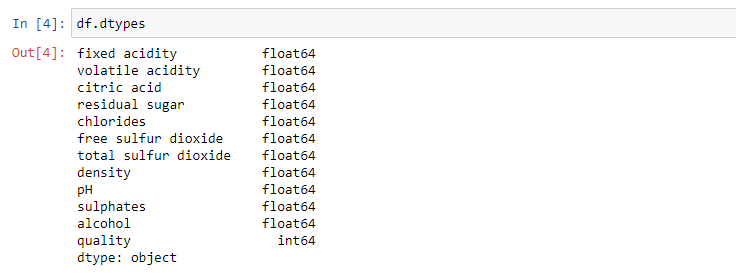
Now I am going to upload or read the files/data-sets using pandas. For this we used read\_csv

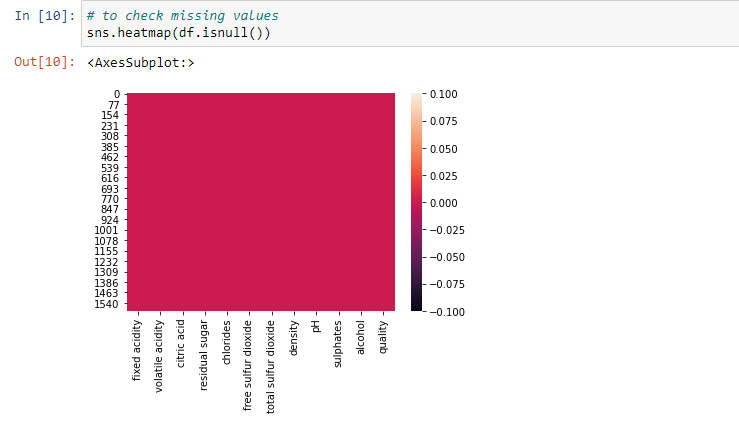
And let me get the top 5 values. We can get using the head function. Hence the code would be df.head().



­

Checking all variable datatype using df.dtypes

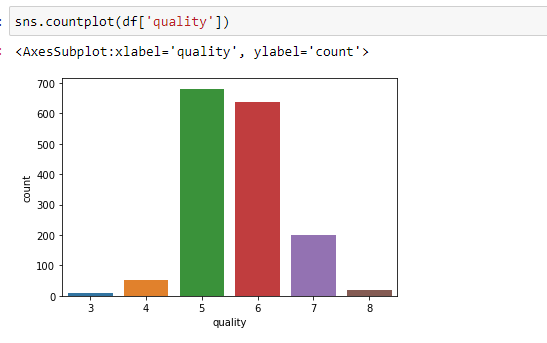




We having all numerical features in data and there is no missing values in data

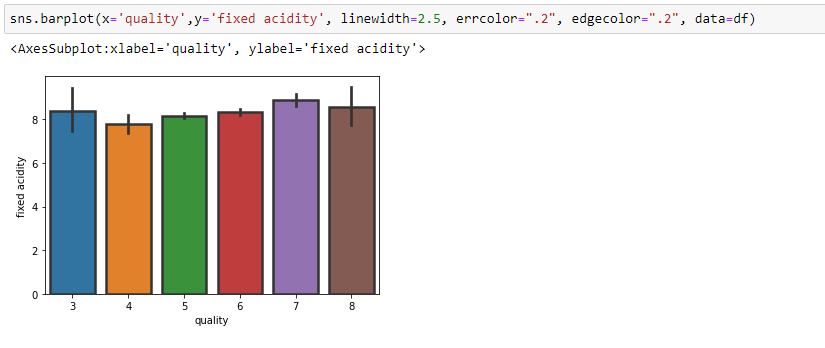
**Exploratory Data Analysis:**

We know that the “image speaks everything” here the visualization came into the work, we use 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.

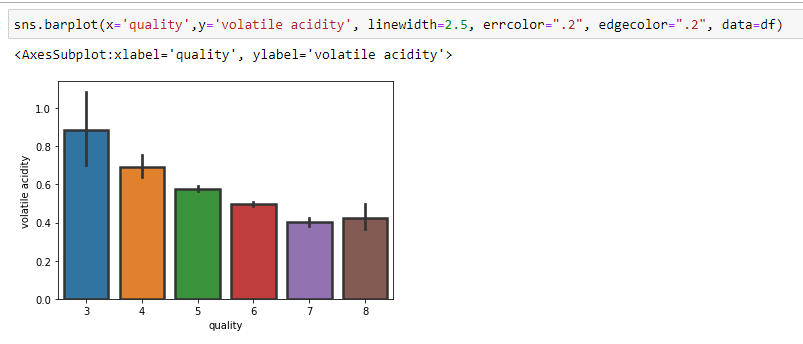


Here we can see most of samples of wine around 600 to 700 relay in no 5 & 6 and around 200 in no. 7 and more which is count as good wine

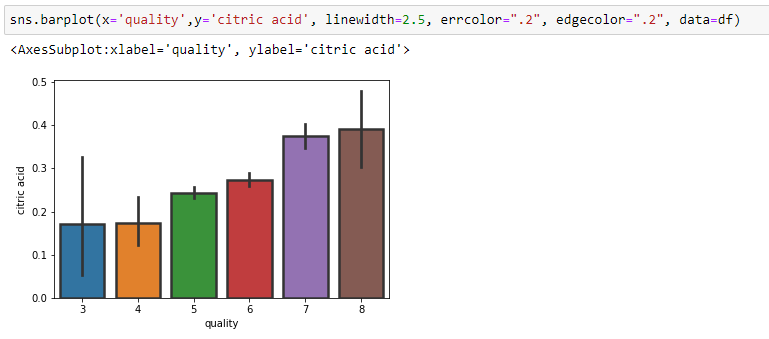
so whole model will bias to these three quality no (will treat in imabalance data)



we can see here fixed acidity is almost common in all the samaples of wine means it doesn't affect the quality of wine that much



In above image we notice less volatile acidity improves the wine quality



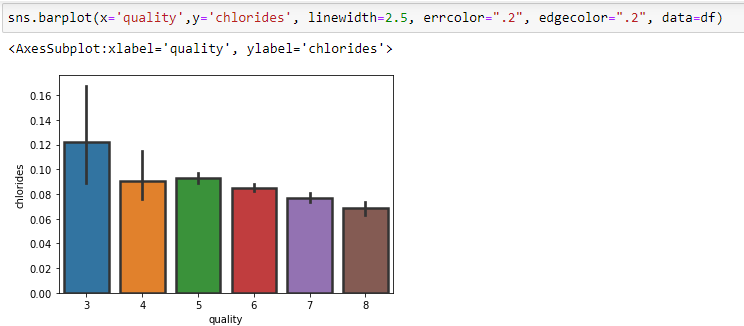
In above image:

Citric acid has a huge impact on quality of wine, increasing of citric acid quantity also improves the quality of wine



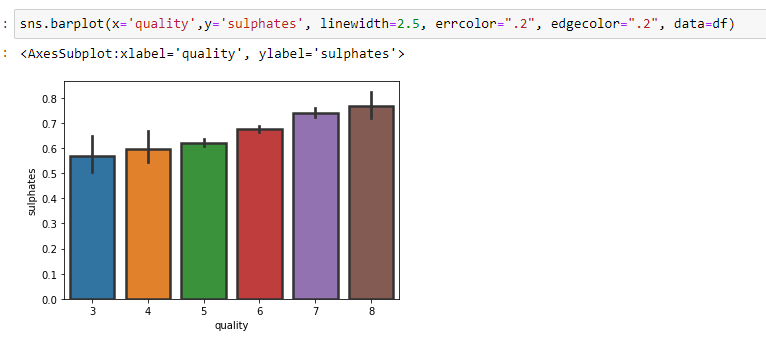
In above picture:

There is no huge impact of residual sugar on quality of wine it's almost common in all qualities.



In above image :

we notice here chlorides has a little impact on quality if it decreases that improves little quality of wine

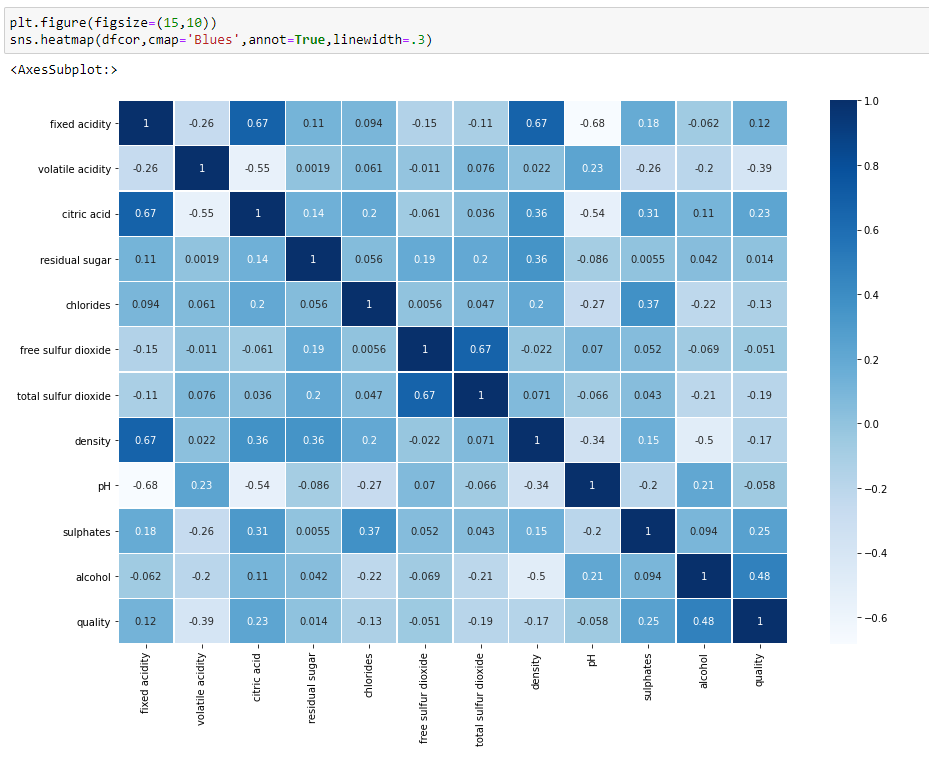


In above image

Here we notice sulphates improves the quality of wine on minor level

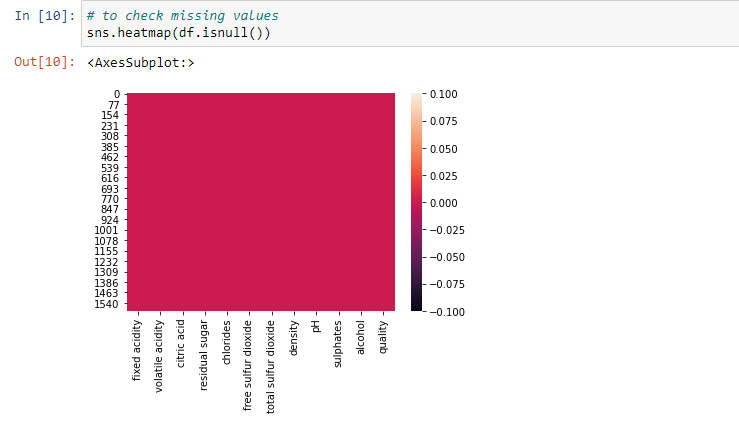
### **Correlation: -**

For checking correlation, we use a statistical method that finds the bonding and relationship between two features.



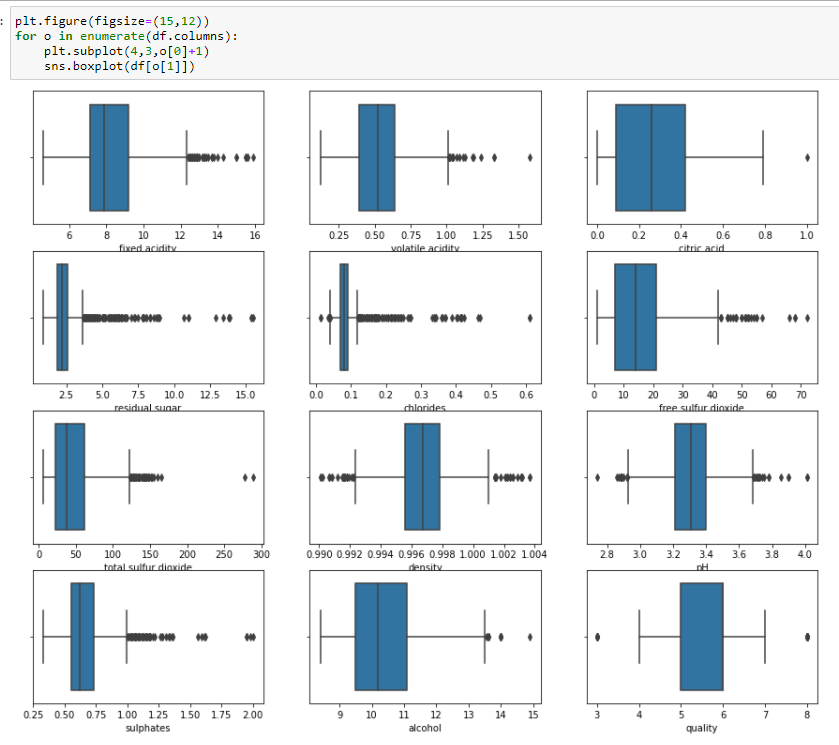
**DATA CLEANING AND STRUCTURING:**

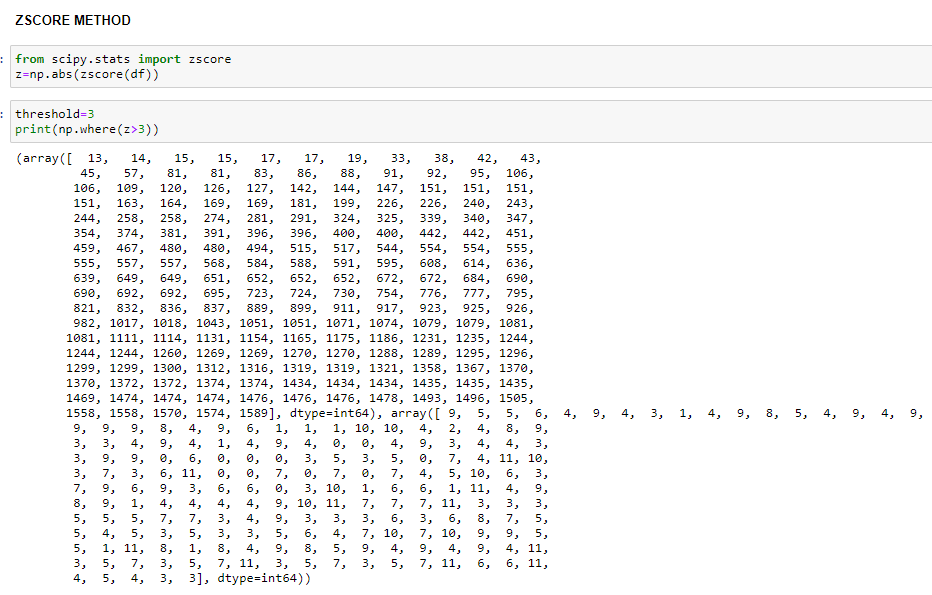
Before we go for modelling the data, we have to check whether the data is cleaned or not. And after cleaning part, we have to structure the Data. For cleaning part, First I have to check whether there exists any missing values. For that I am using the code snippet isnull()



There are no missing values in data so we can move further.

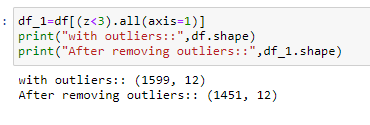
**Outliers Handling:**



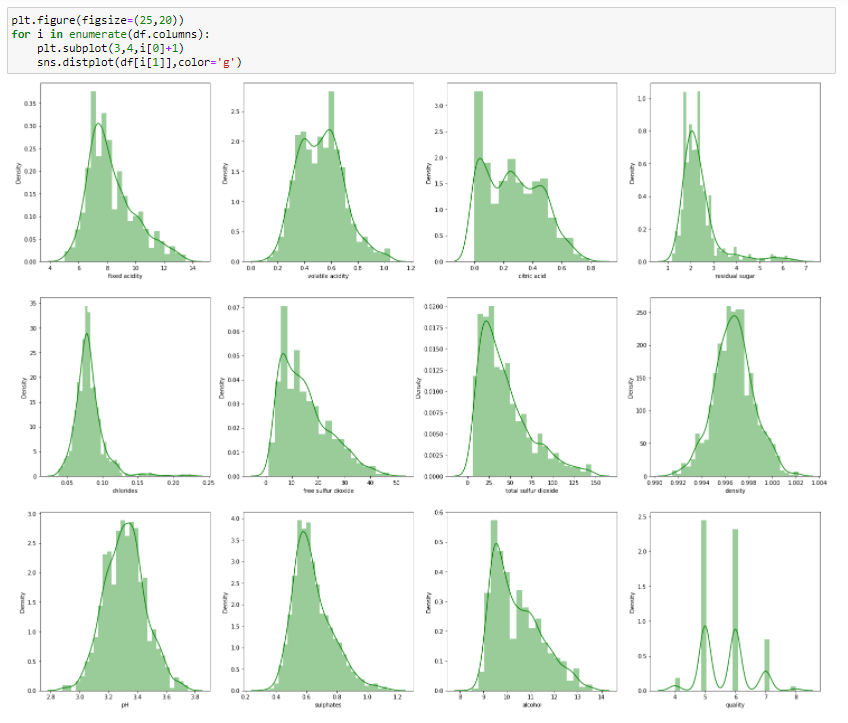


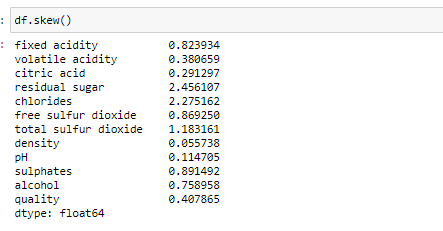
In above image we notice there are some outliers in dataset

Have used **ZSCORE** to remove outliers



**Skewness Handling:**





In above image we can notice there is skewness in data, skewness more than +/- 0.5 have been handled

#### These are the features I have removed skewness from

**fixed acidity**

**residual sugar**

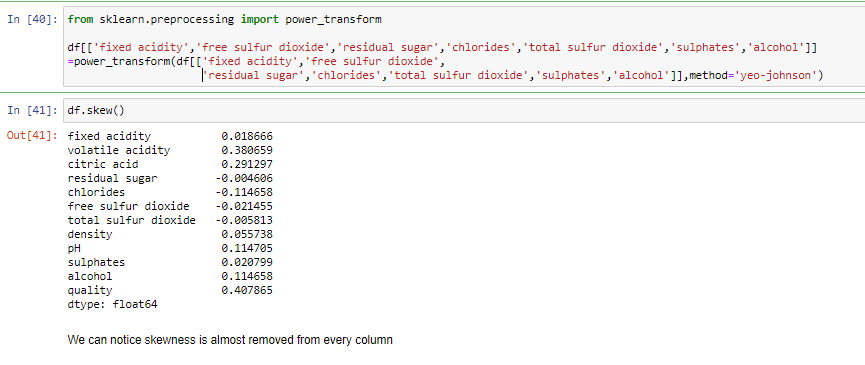
**chlorides**

**free sulfur dioxide**

**total sulfur dioxide**

**sulphates**

**alcohol**

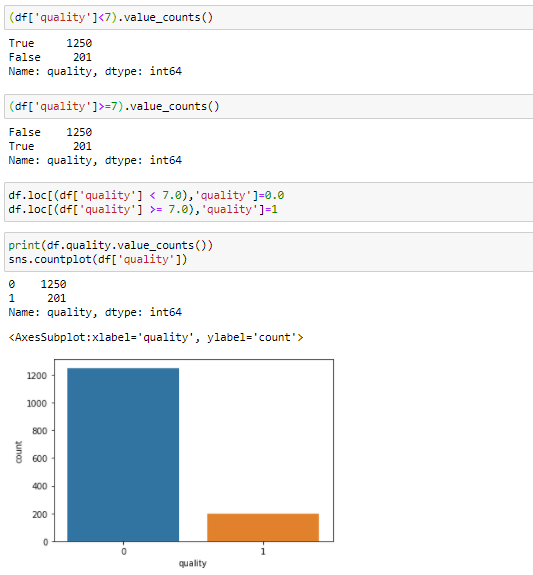


I have used power transform method to handle skewness from data

In above image, can be notice skewness is almost removed.

**Converted target into two categories good and not good**

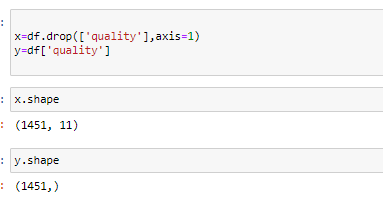
**7** or higher getting classified as **'good/1'** and the remainder as **'not good/0'**



Here we can see our target variable (quality)converted into two categories **0 & 1**

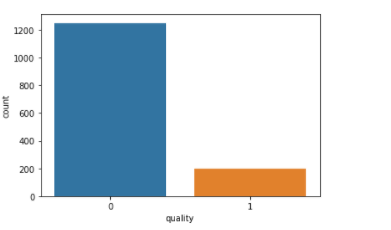
After converting the data, we noticed that there is imbalance in quality lets handle that.

**First dividing the data into two parts x and y**

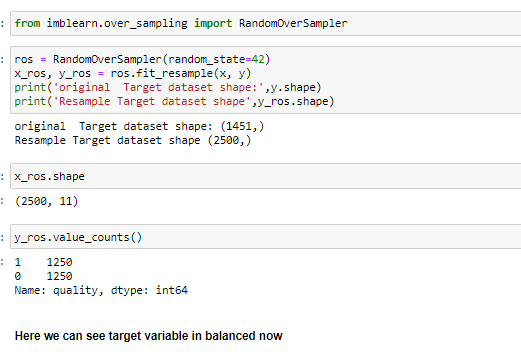


**Imbalance learn:**

we have very less amount of data in good quality of wine that need to be balance for the better modelling and better prediction

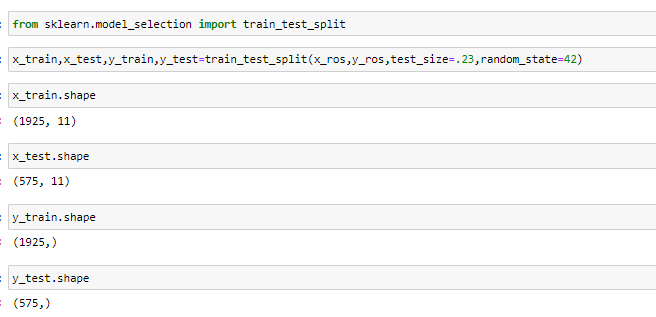


According to data I am going to Using RandomOverSampler because there is very less data



Now we have balance data for model

***Spliting the data:***

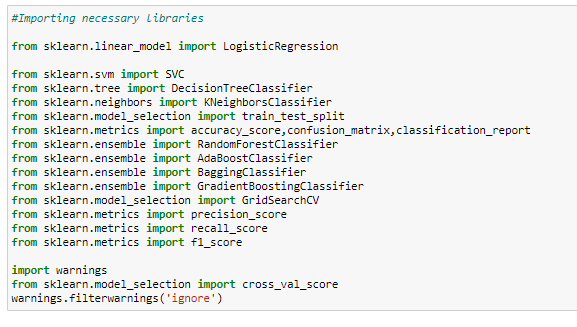


***Building Machine Learning Models***:

***SAMPLING TECHNIQUES AND NEED FOR THAT*** :

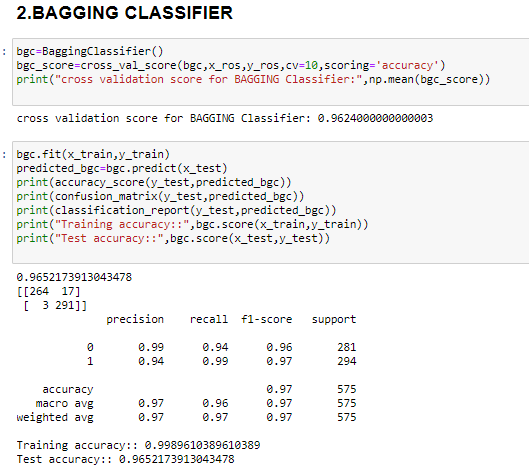
There are many sampling techniques like Random Sampling, Stratified Sampling etc. The major purpose is to improve the accuracy which can be obtained by hiding some portion of train data and running the model so that on an average the one that gives higher accuracy can be taken for test data.

Importing some import libraries



**Bagging Classifier:**

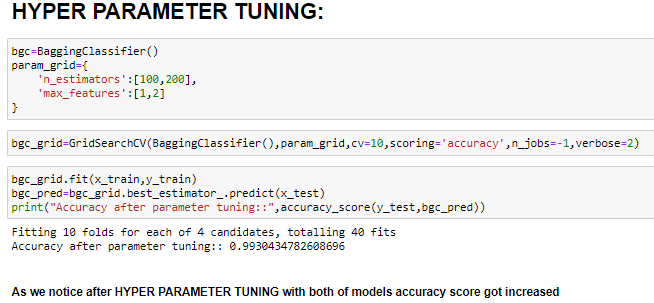
I have tried various techniques like Random Forest, Ada Boost, Decision Tree, SVR etc. and came to conclusion that the above code gave maximum accuracy.



# HYPER PARAMETER TUNING:

To enhance the accuracy tuning the model .

I have used the **GridSearchCV** grid for tuning the model so we can get the best parameters



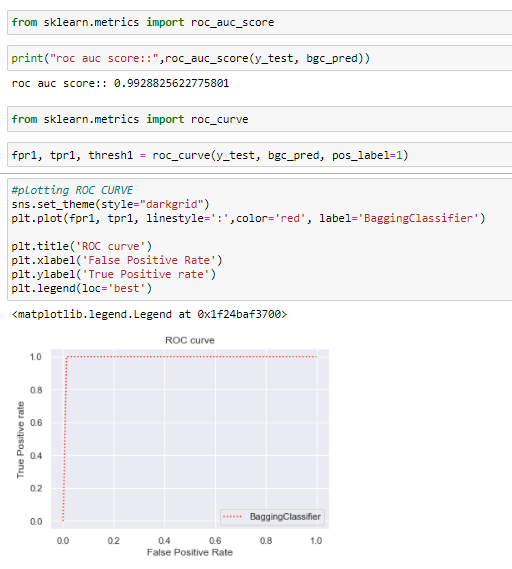
After tuning the model, we got improvement in data accuracy

As we can see average accuracy is **99.30%** . I have tried with other parameters as well but this the best I got.

**Classification Report** :



**auc\_roc\_score: 99.28%**



***CONCLUSION*:**

Key Findings and Conclusions of the Study:

* 1. Volatile Acidity has a high impact on red wine quality, for best quality of wine volatile acidity quantity should be less
  2. citric acid has a huge impact on quality of wine, if we increase quantity of citric acid that also improves the quality of wine
  3. chlorides have a little impact on quality if its quantity decrease that improves little of wine quality
  4. high quantity sulphates also improve the quality of wine