**Introduction**

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:

## 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.

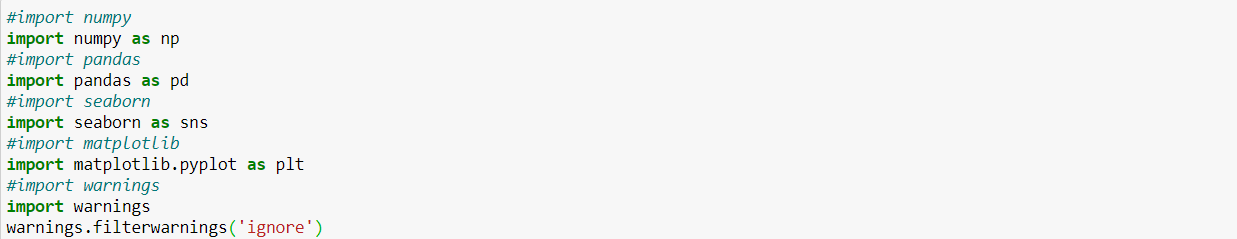
* **volatile acidity :**   Volatile acidity*is the*gaseous acids present in wine.
* **fixed acidity :**Primary **fixed acids** found in wine are **tartaric**, **succinic**, **citric**, and **malic**
* **residual sugar :**Amount of sugar left after fermentation.
* **citric acid :** It is weak organic 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.

For the next step we have to import some important library :

**Importing modules**

Let’s import,

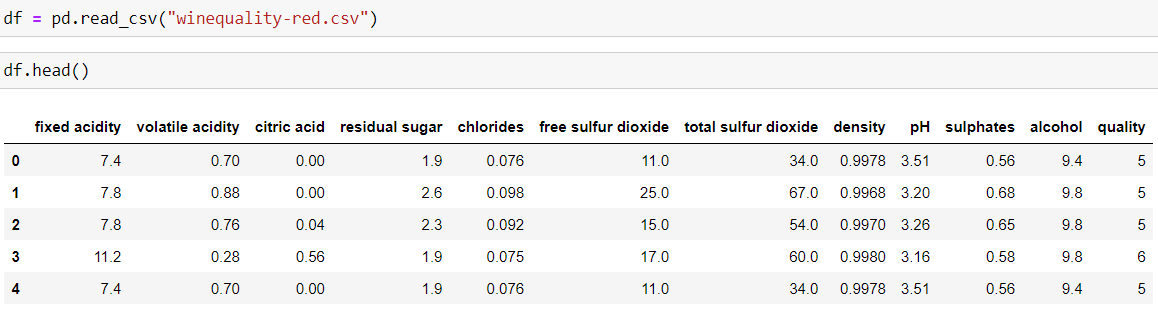


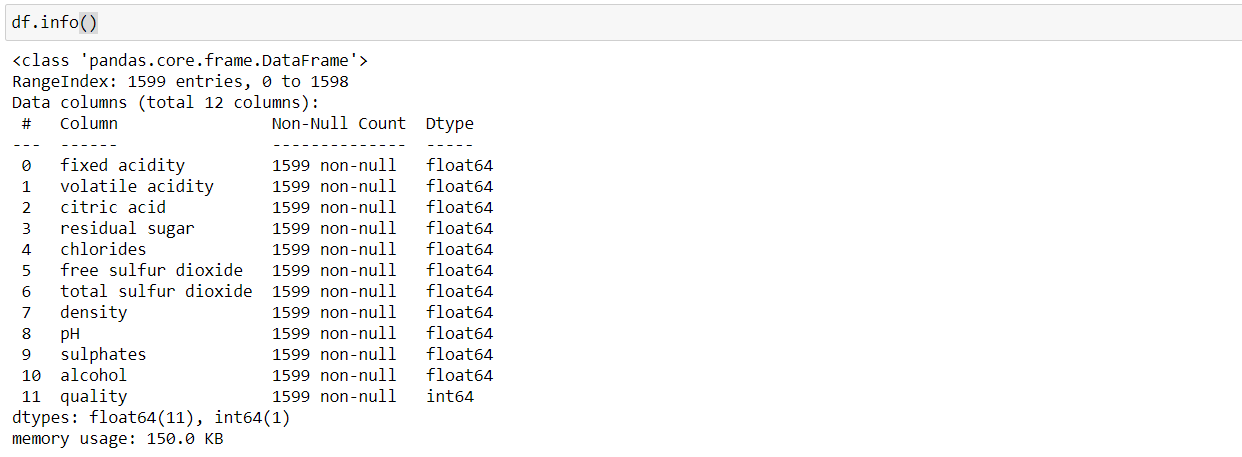
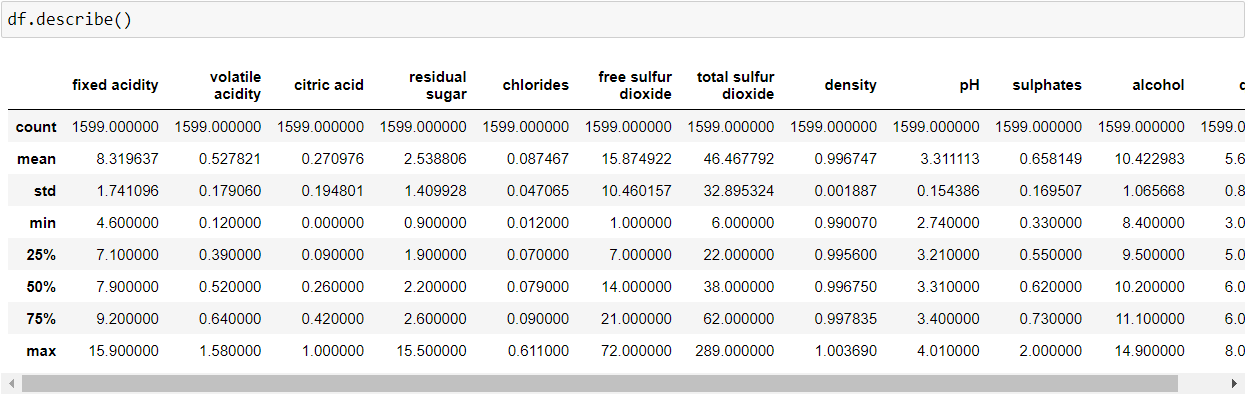
Let’s we take brief about these libraries, **pandas** are used for data analysis **NumPy** is for n-dimensional array **seaborn** and **matplotlib** both have similar functionalities which are used for visualization and warnings are used to ignore all the warnings that comes while running the code.

The next step is to read the wine quality dataset and see their information:

**Study dataset**

For the next step, we have to check what technical information contained in the data,

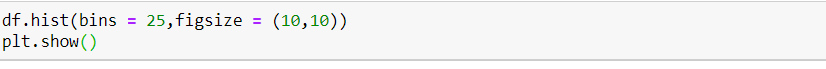


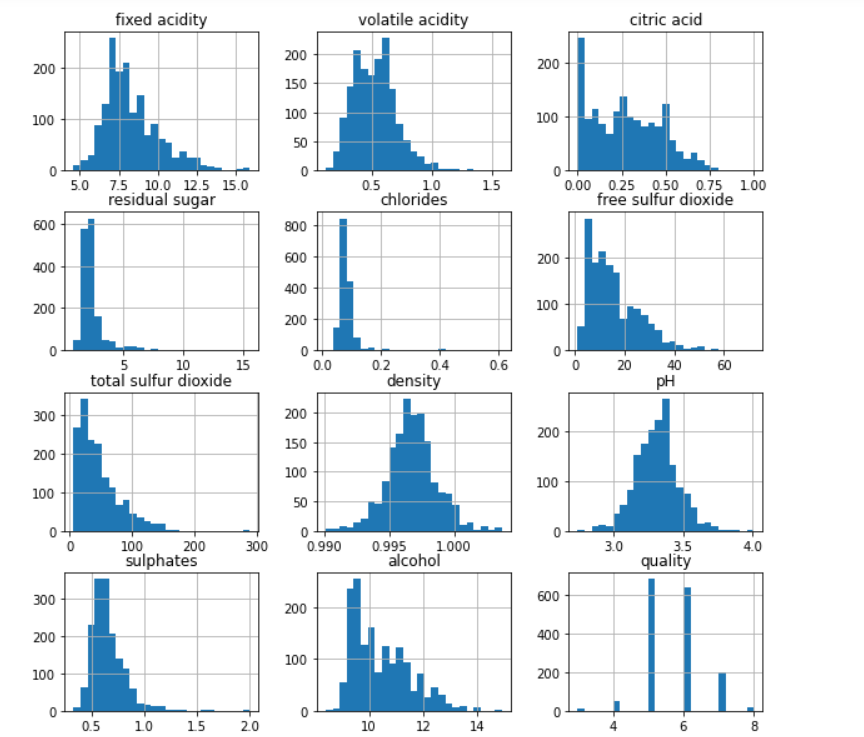


As we see in the above image, there is vital information on features and with this information, we will process our next work.

**Visualization**

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.



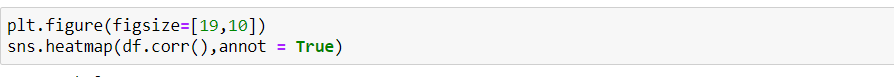


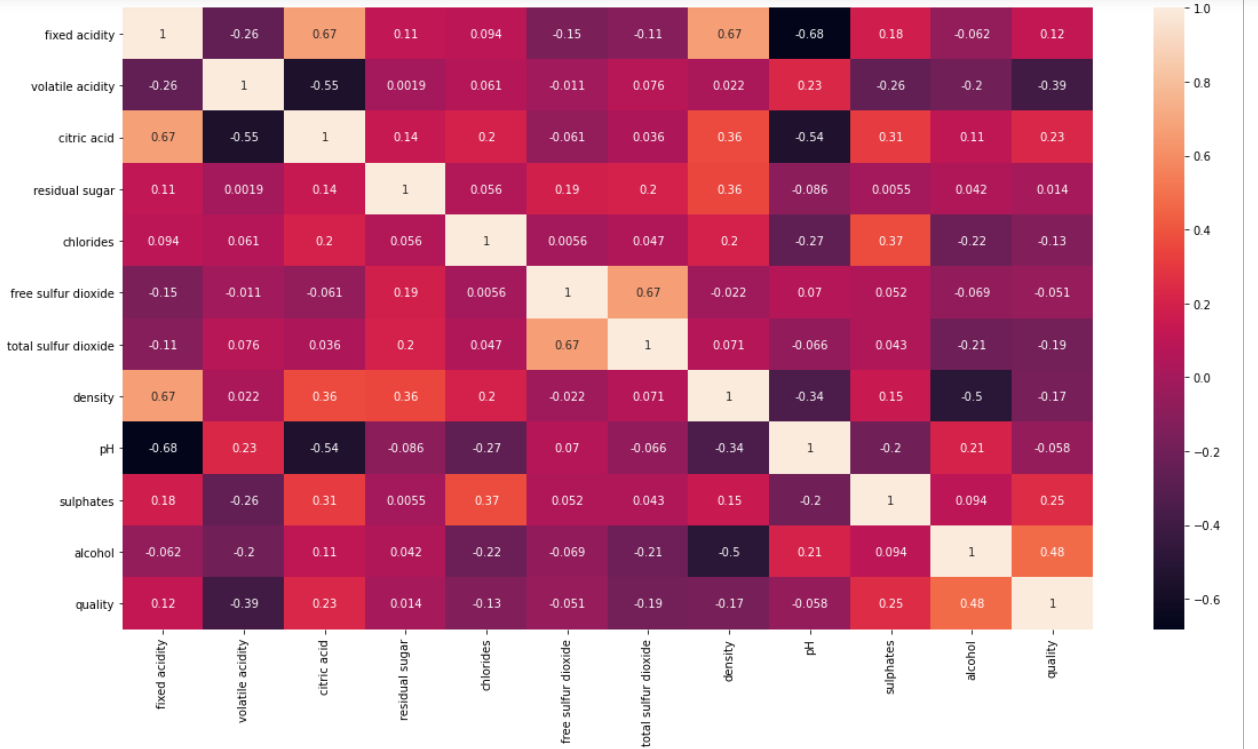
The above image reveals that how that data is easily distributed on features.

When we performing any machine learning operations then we have to study the data features deep, there are many ways by which we can differentiate each of the features easily. Now, we will perform a correlation on the data to see how many features are there they correlated to each other.

**Correlation:-**

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

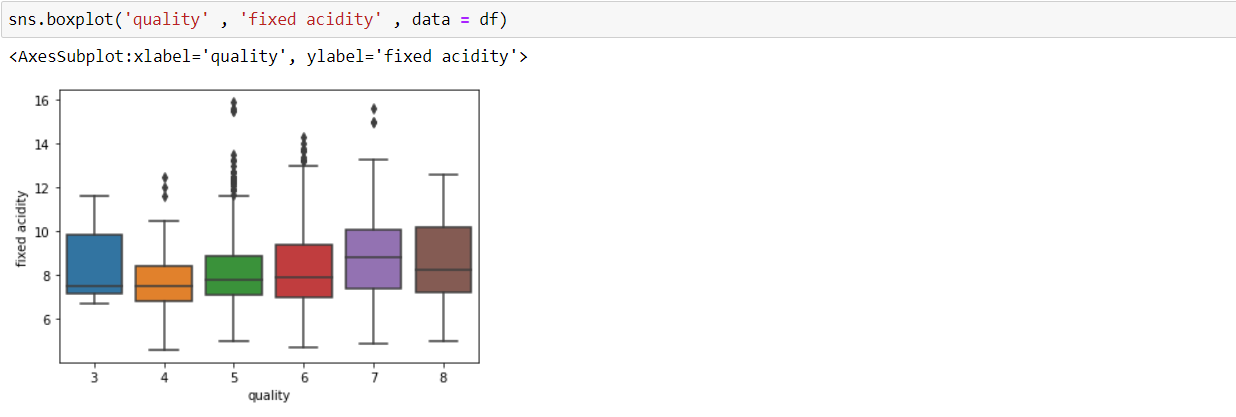
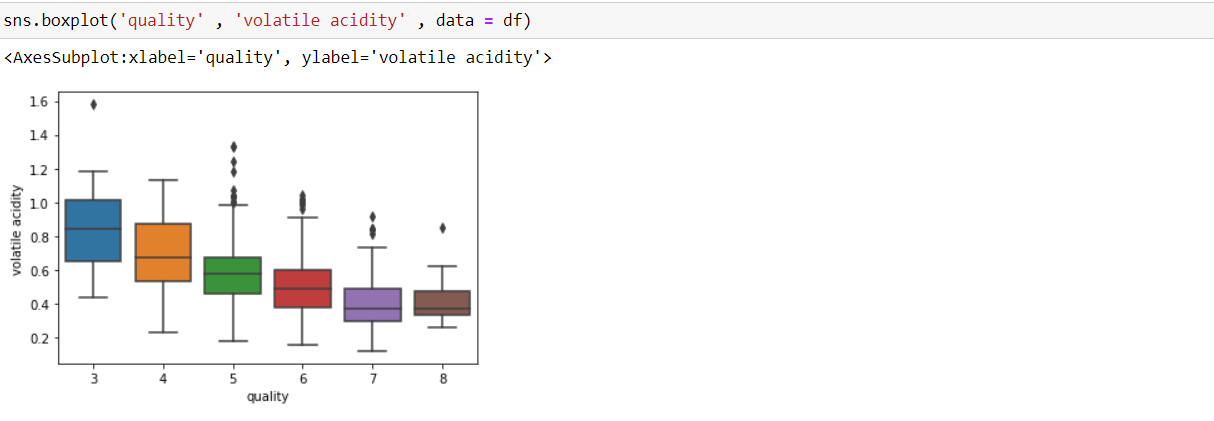
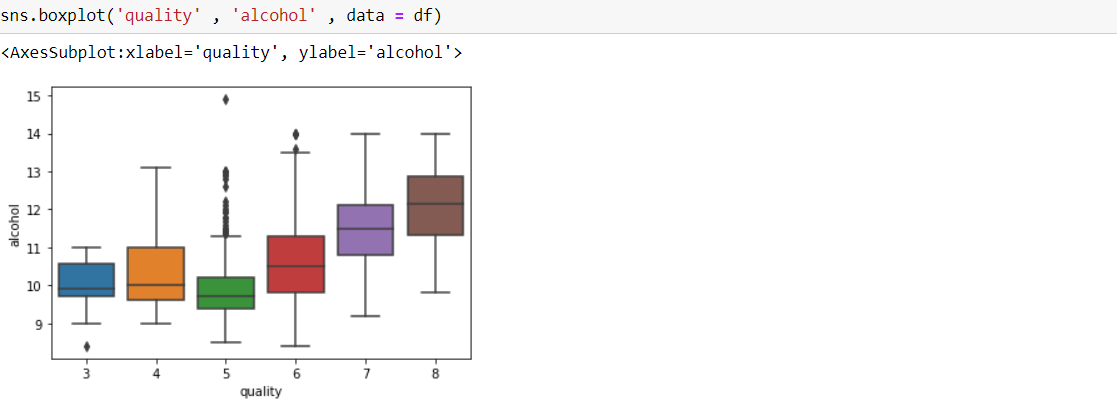
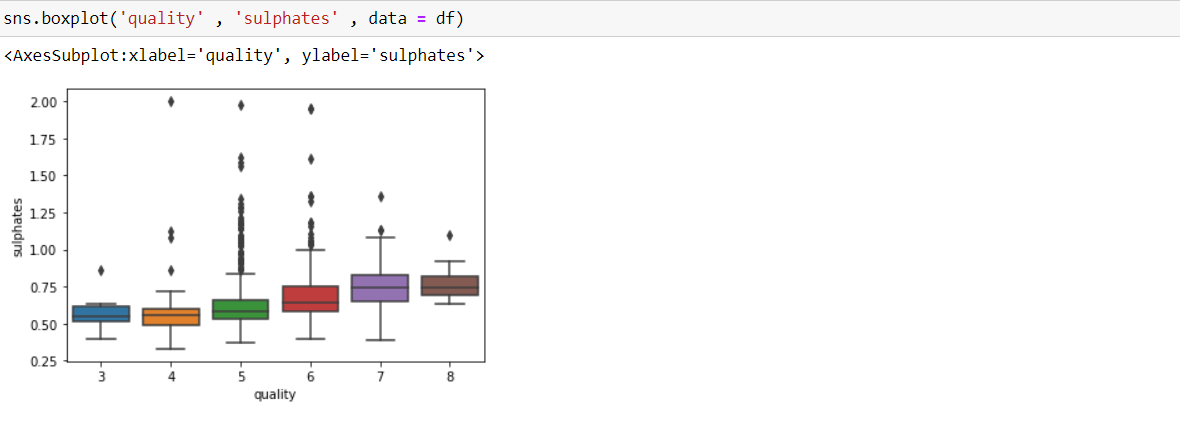
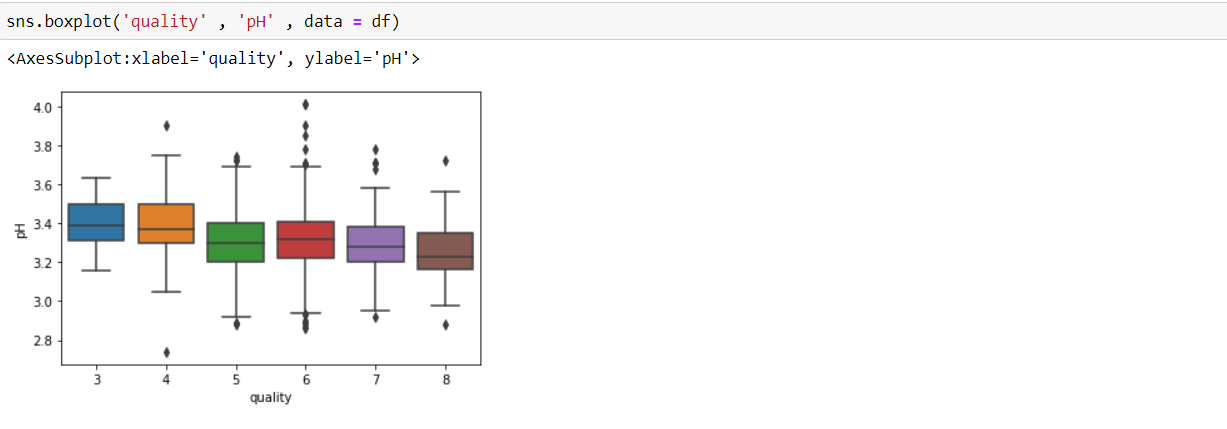
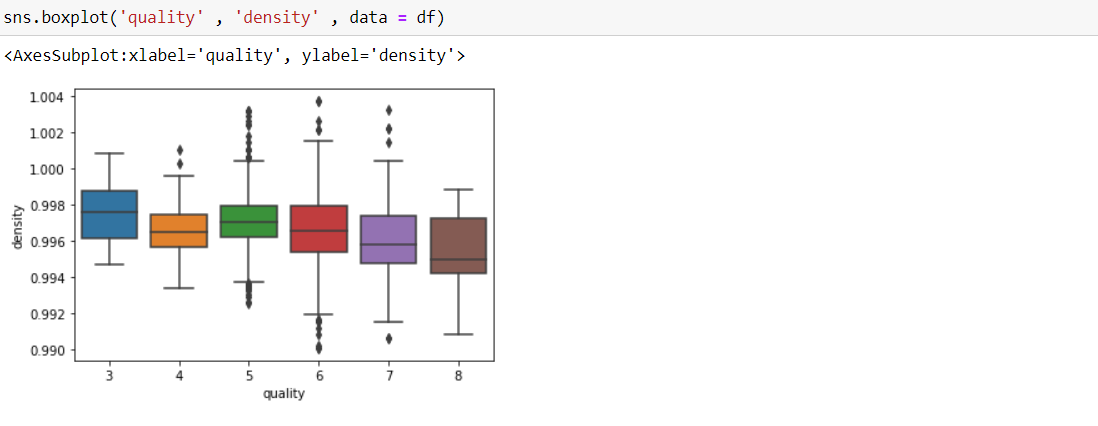
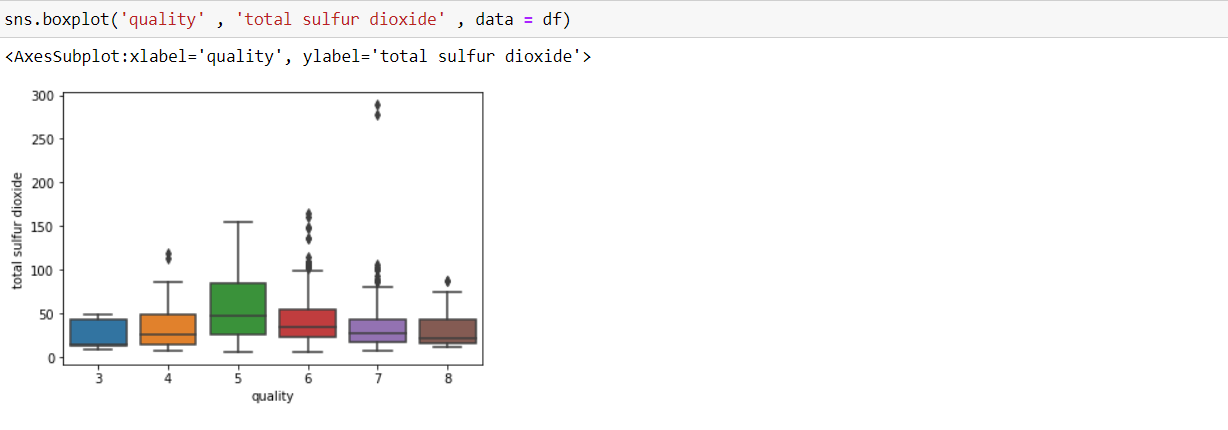
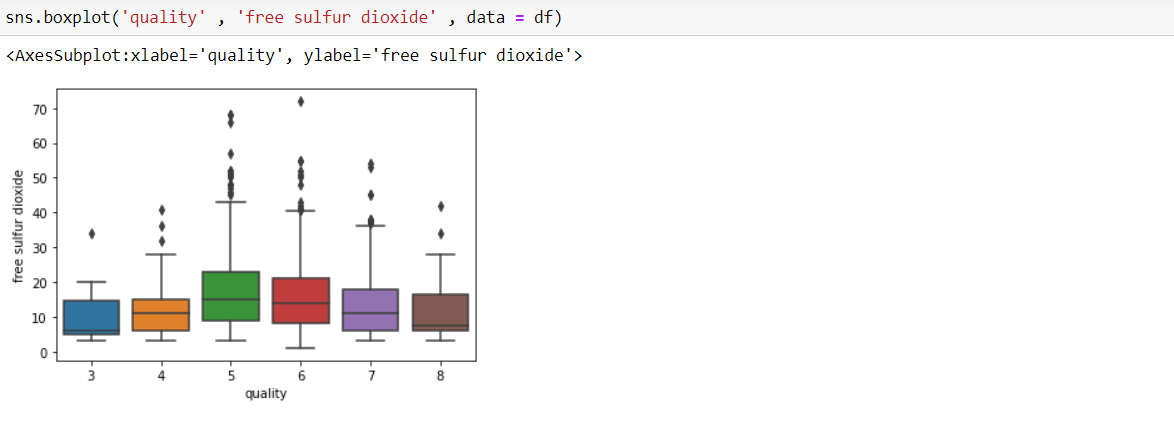
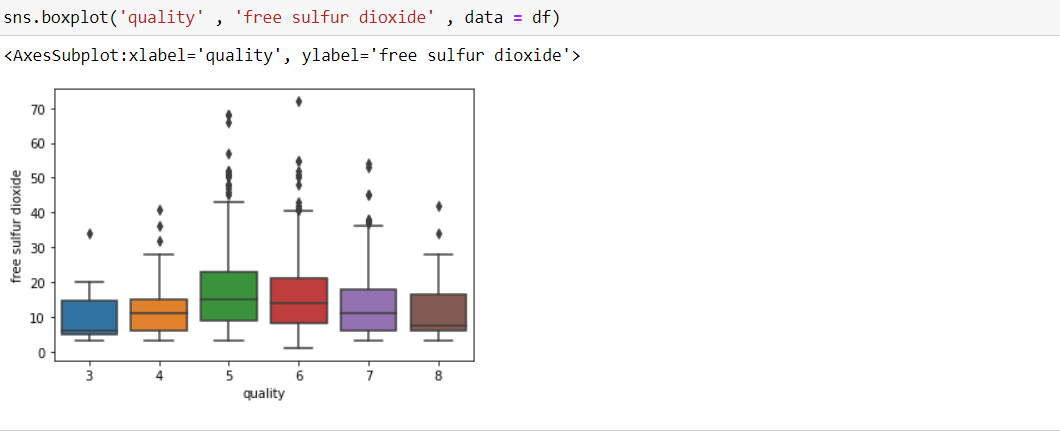
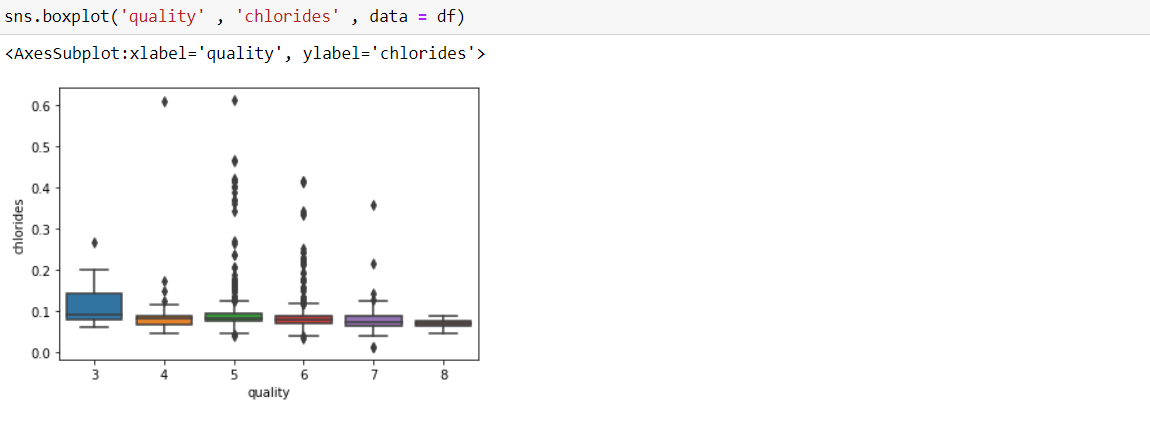
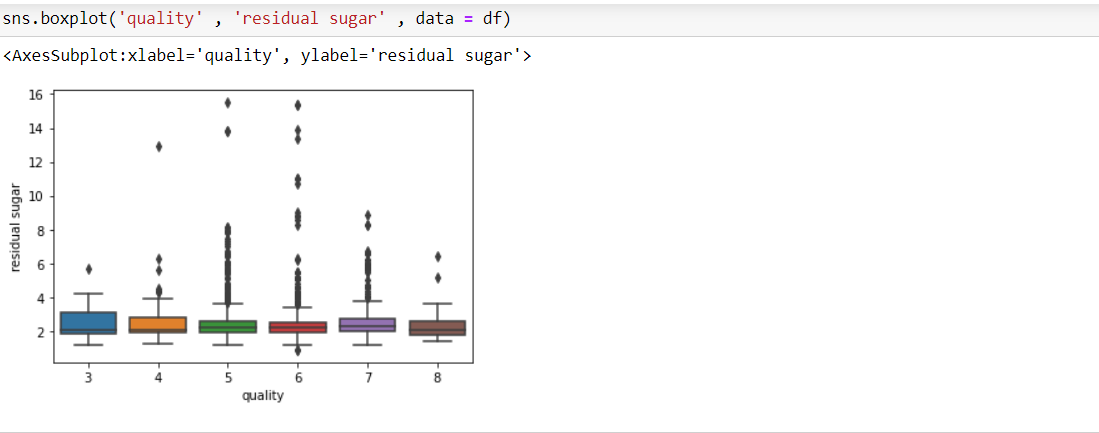
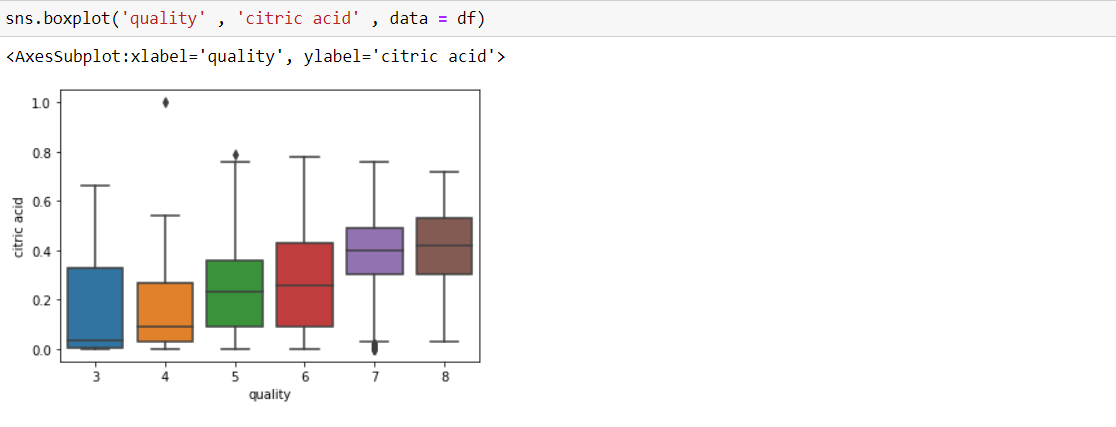




Using the count plot we will count the target variables.

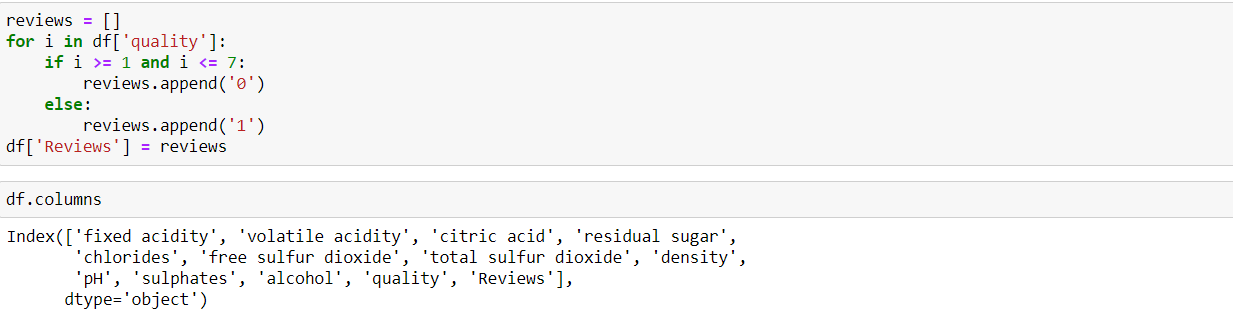


Now, for further we must have to know the outliers from the data so that we can clean the data as to get the better results for prediction. So to check the outliers we used box plot. Let’s see the outliers now:-

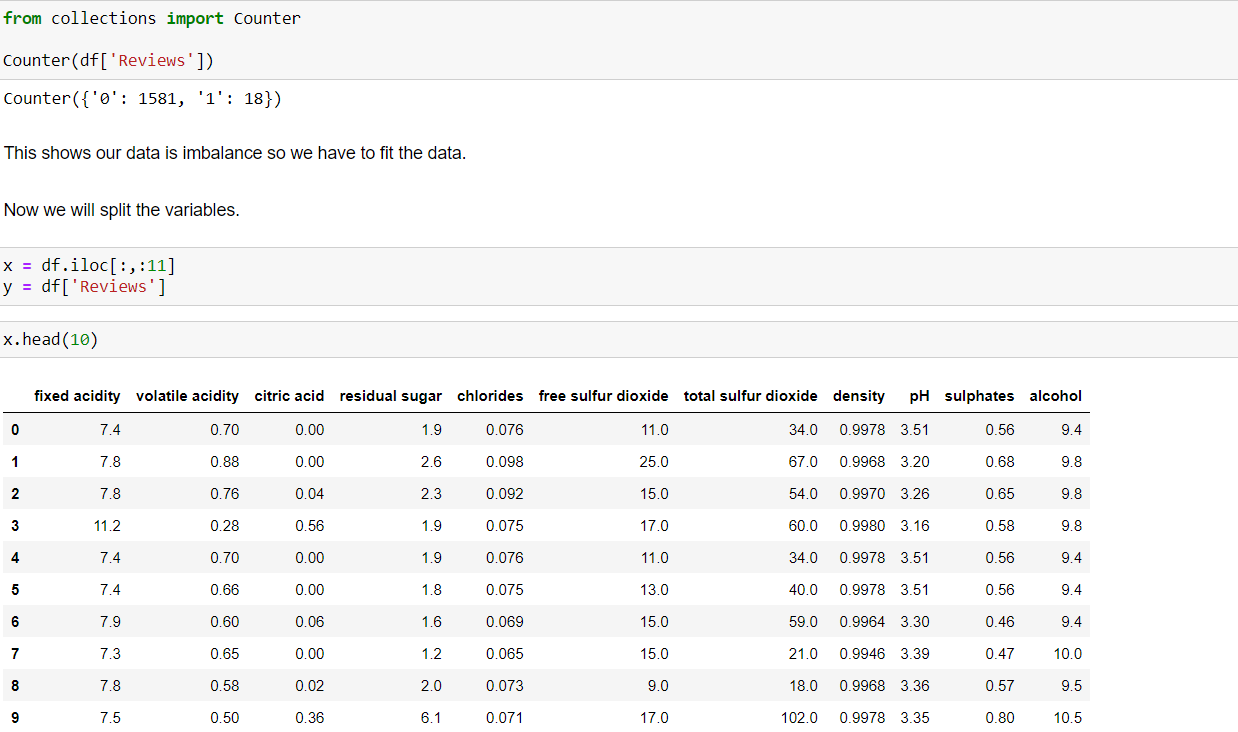
 

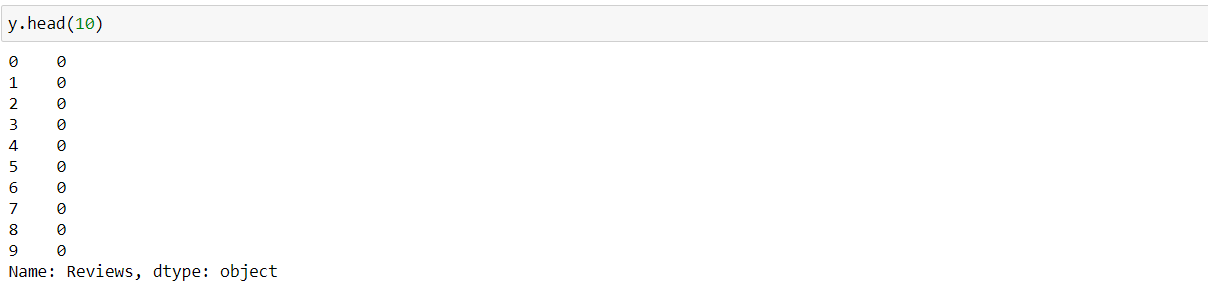
In above graph we have checked the outliers using the dependent variables and target variable. We observed that every plot is containing black dots that are outliers. We will take out the outliers are it can make our accuracy even more better.

Now we will create a new column to check the reviews and count the data using counter and try to balance the data as working on imbalance data we will not get the desired results.



Above we have used df.columns to check whether the column is created or not.

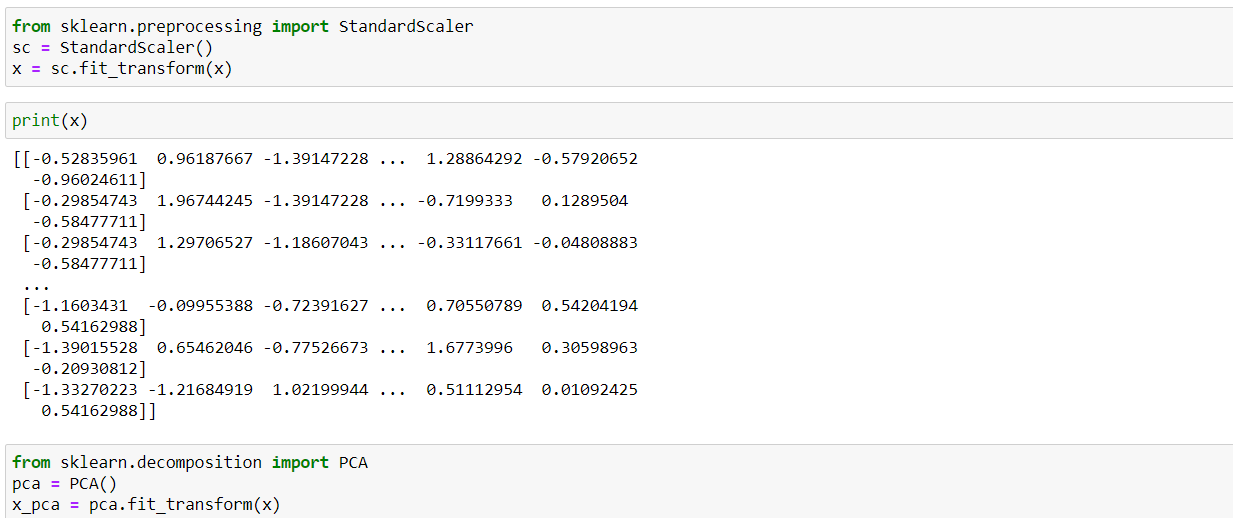




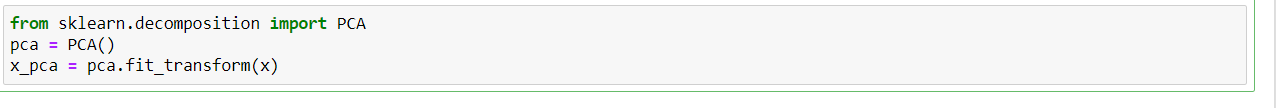
From above we can see that the data is imbalance. We have to balance the data using standard scaler PCA.

Standard Scaler

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



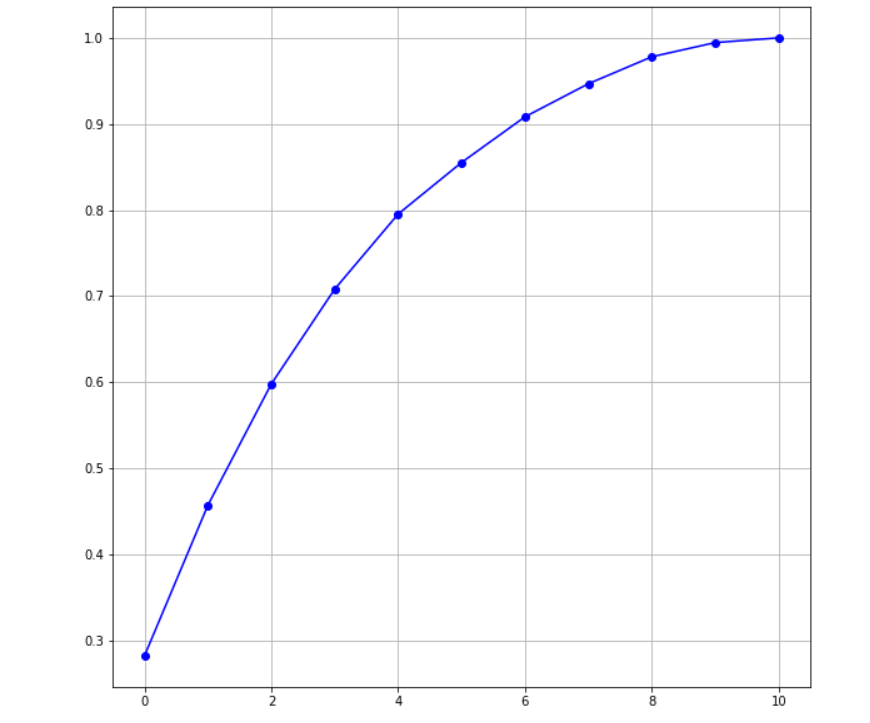
Now we will plot the graph to see the principle components.



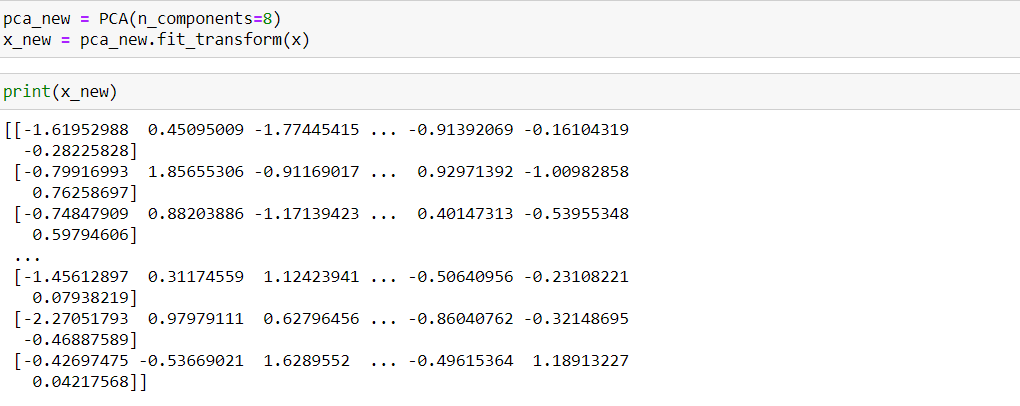
PCA

An important machine learning method for dimensionality reduction is called Principal Component Analysis.

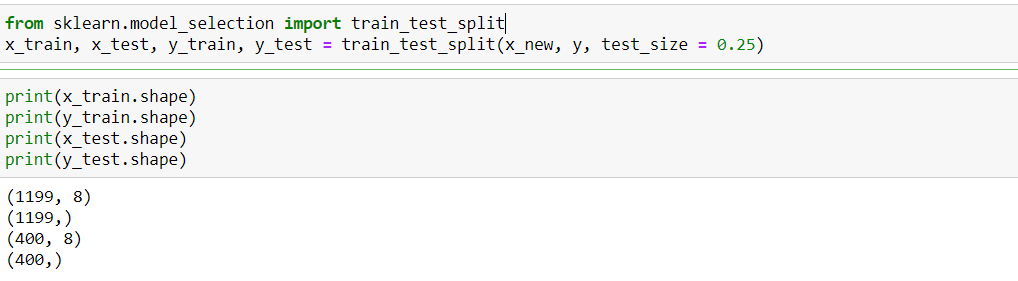
It is a method that uses simple matrix operations from [linear algebra](https://machinelearningmastery.com/gentle-introduction-linear-algebra/) and statistics to calculate a projection of the original data into the same number or fewer dimensions.



Looking at graph we can observe that 8 principal components attribute have 90% of variation in the data.



Next we will split the data into a training and test set so that I could cross-validate my models and determine their effectiveness.



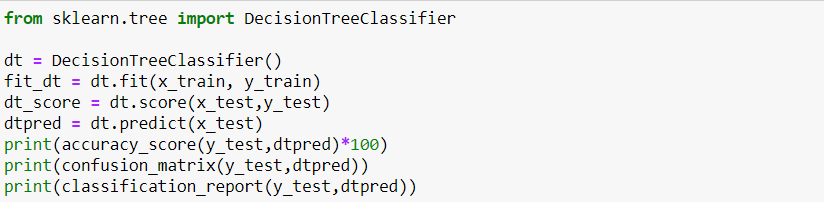
**Modelling**

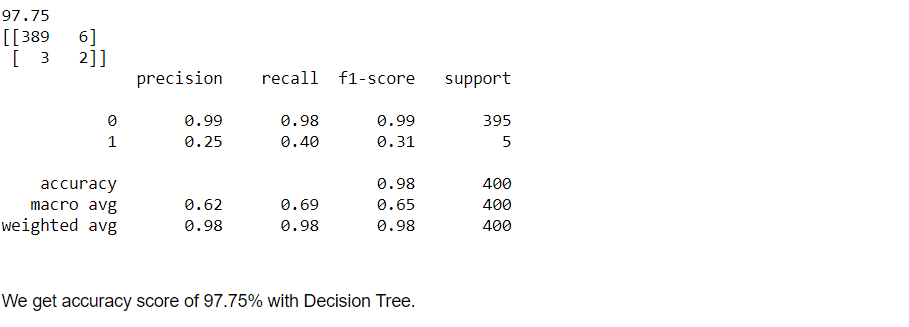
For this project, We wanted to compare four different machine learning models: Decision trees, Random forests, Support Vector Machine, Logistic Regression . For the purpose of this project, we wanted to compare these models by their accuracy.

## Model 1: Decision Tree

## 

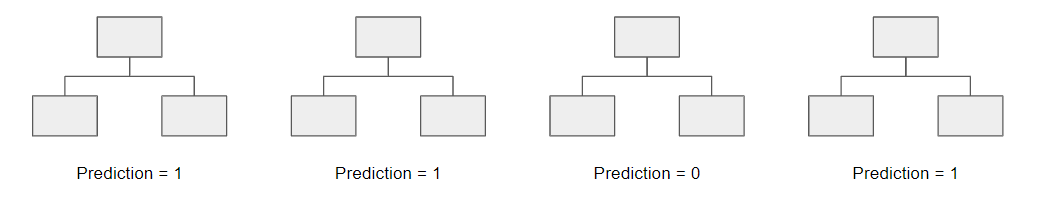
## Decision trees are a popular model, used in operations research, strategic planning, and machine learning. Each square above is called a node, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the leaves of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.

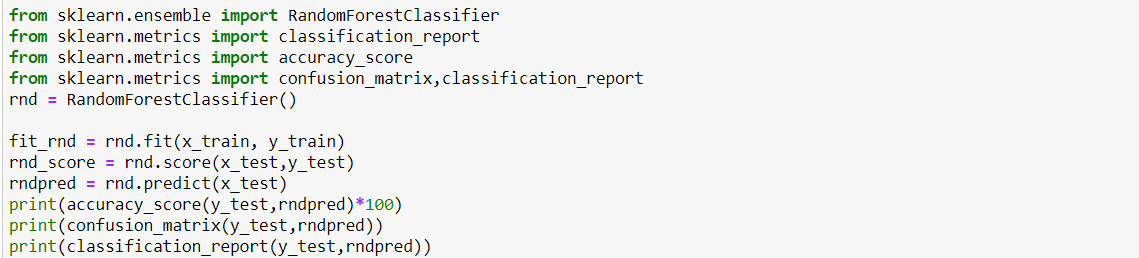


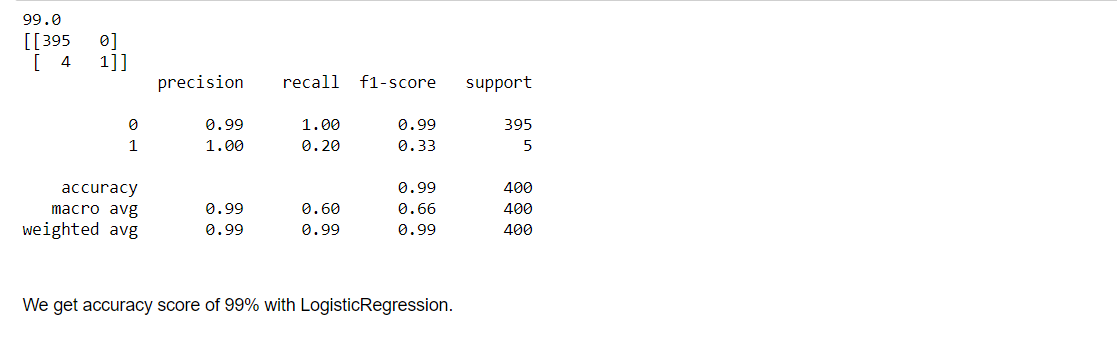


## Model 2: Random Forest

## Random forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) technique that builds off of decision trees. Random forests involve creating multiple decision trees using [bootstrapped datasets](https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/) of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. What’s the point of this? By relying on a “majority wins” model, it reduces the risk of error from an individual tree.

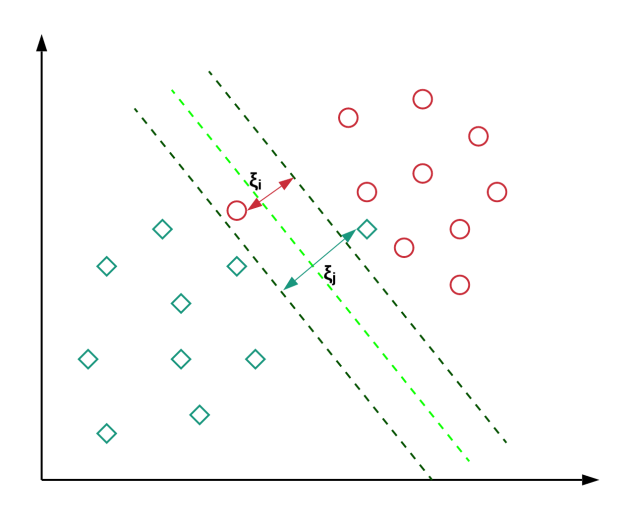


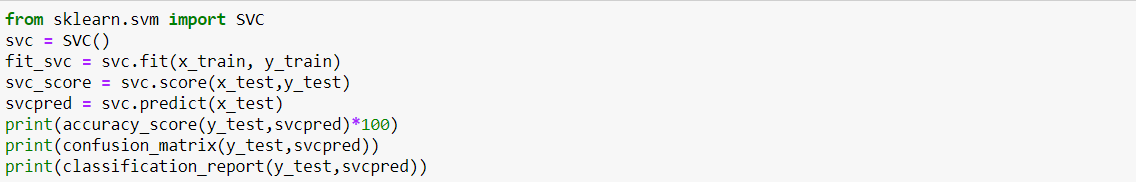


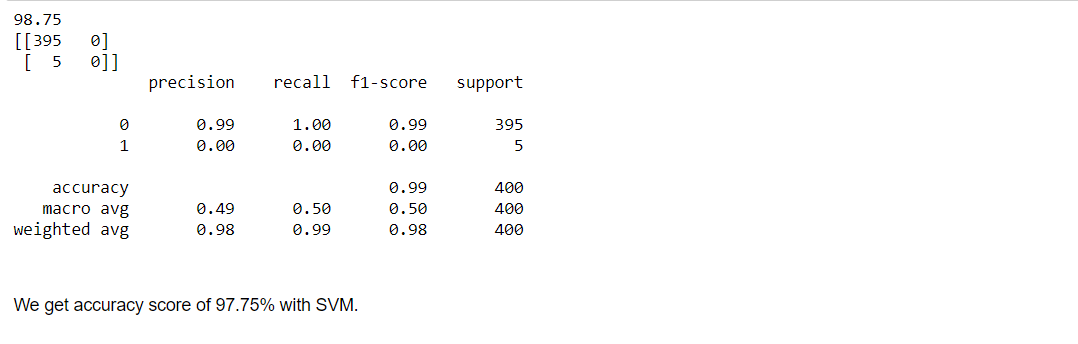


## Model 3. Support Vector Machine

## “Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) that can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate.







## Model 4. Logistic Regression

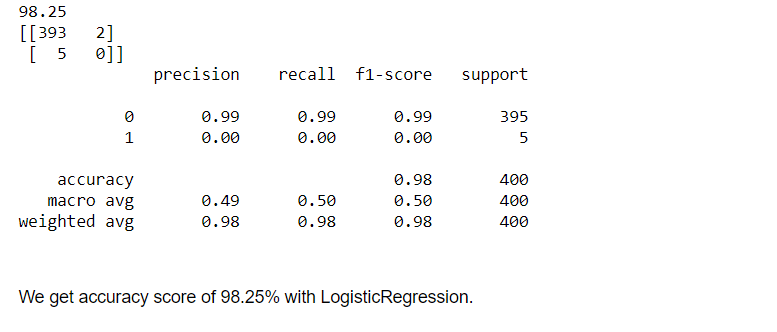
Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

## 

## 



Now, as we get the accuracies of all the models. We have to find the best model.

Here we have used the cross validation score to find the best model.



Finally we get the best model i.e Random Forest Classifier. With the help of we will now find the best params from hyper parameter tuning using grid cv search.

Hyper Parameter Tuning

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

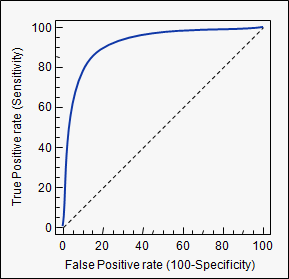


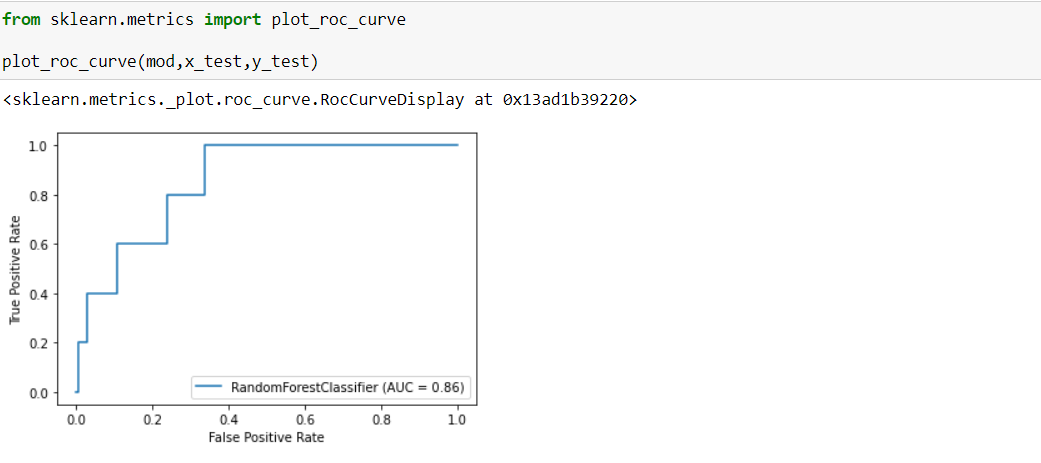
ROC Curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

* True Positive Rate
* False Positive Rate

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.





**By looking into the details, we can see that good quality wines have higher levels of alcohol on average, have a lower volatile acidity on average, higher levels of sulphates on average, and higher levels of residual sugar on average.**