**Red Wine Quality Prediction by using Machine Leaning Algorithm**

**Problem Statement:**

**Wine** is an [alcoholic drink](https://en.wikipedia.org/wiki/Alcoholic_drink) typically made from [fermented](https://en.wikipedia.org/wiki/Fermentation_in_winemaking) [grapes](https://en.wikipedia.org/wiki/Grape). [Yeast](https://en.wikipedia.org/wiki/Yeast_in_winemaking) consumes the [sugar](https://en.wikipedia.org/wiki/Sugar) in the [grapes](https://en.wikipedia.org/wiki/Grape) and converts it to [ethanol](https://en.wikipedia.org/wiki/Ethanol) and [carbon dioxide](https://en.wikipedia.org/wiki/Carbon_dioxide), releasing [heat](https://en.wikipedia.org/wiki/Heat) in the process. Different varieties of grapes and [strains](https://en.wikipedia.org/wiki/Strain_(biology)) of yeasts are major factors in different styles of wine.

What if you think about the quality of wine, how can you differentiate the wine according to their quality?

According to experts, the wine is differentiated according to its **smell**, **flavour**, and **colour,**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:

**Data Analysis.**

In 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.

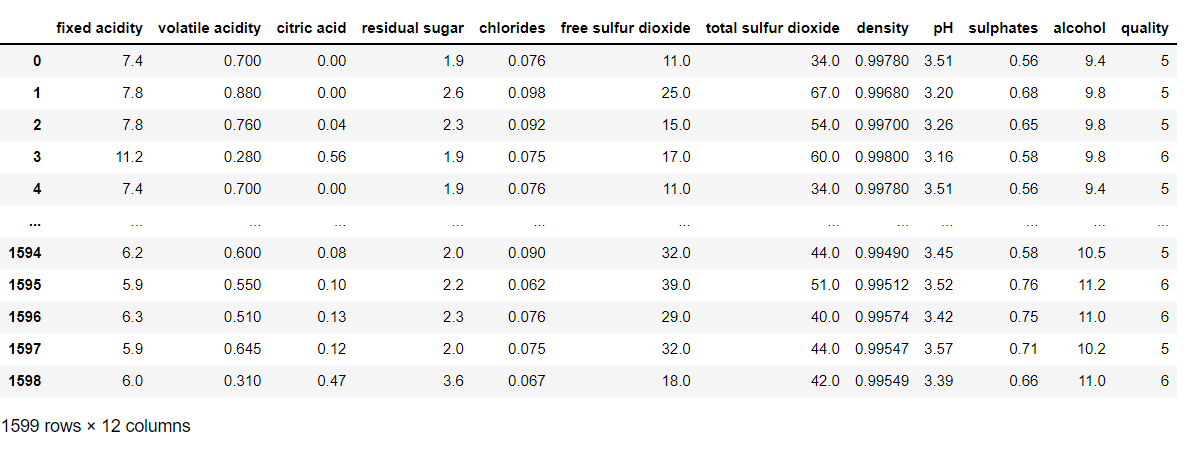
Rather than chemical features, you can see that there is one feature named **Type**it contains the types of wine we here discuss on **red** and **white** wine, the percent of red wine is greater than white.

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

Importing modules

Let’s 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.

check what technical information contained in the data,

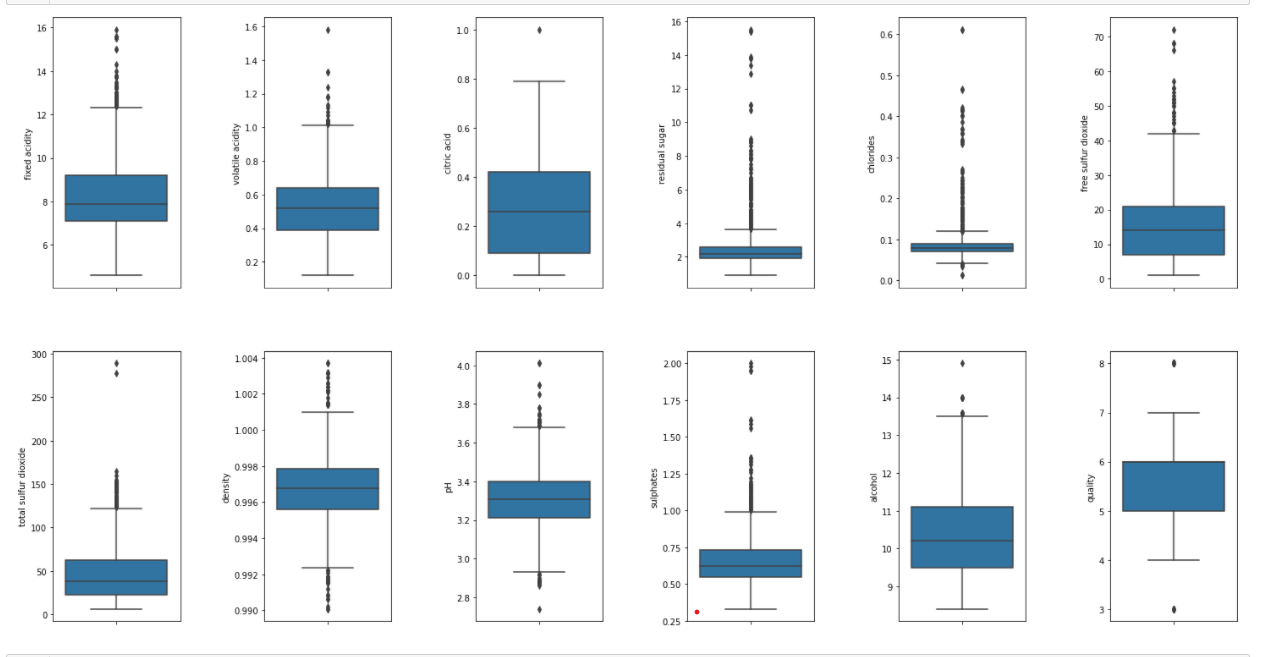


As we see in the above image, there is vital information on features and with this information, we will process our next work.

**EDA Concluding Remark.**

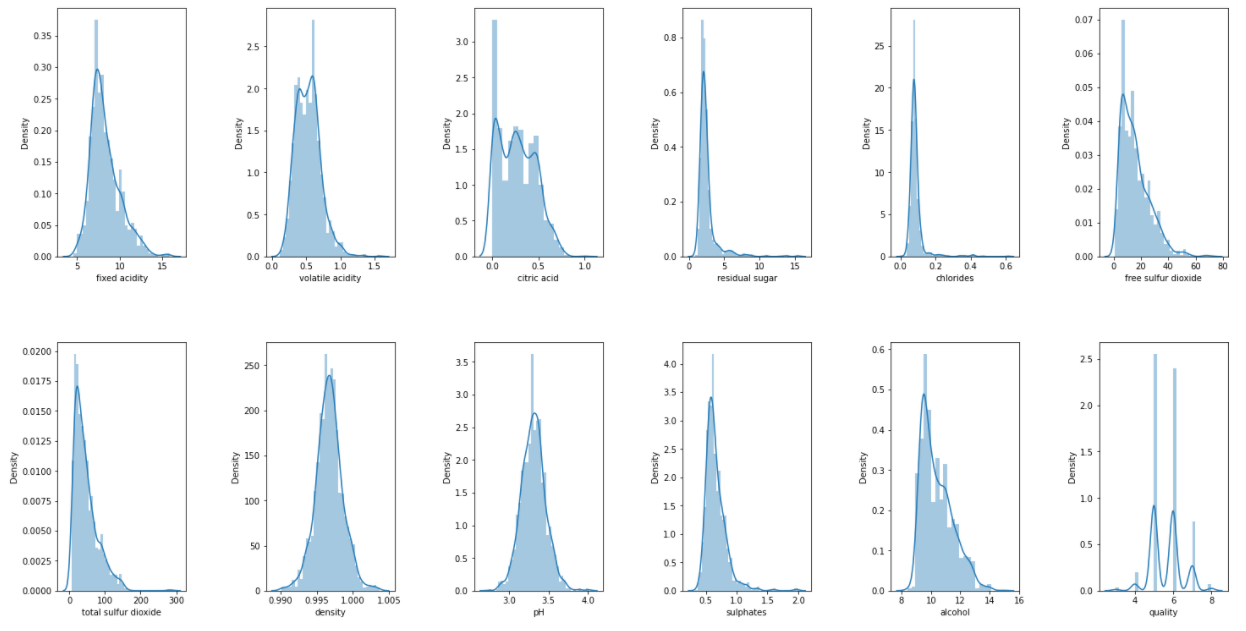
Visualization

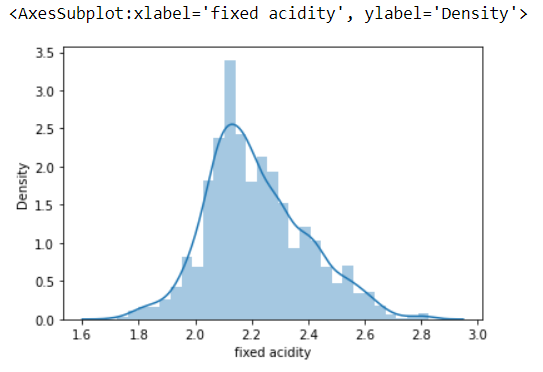
We know that the “image speaks everything” here the visualization came into the work, we use visualization for explaining the data. In other words, we can say that it is a graphic representation of data that is used to find useful information.

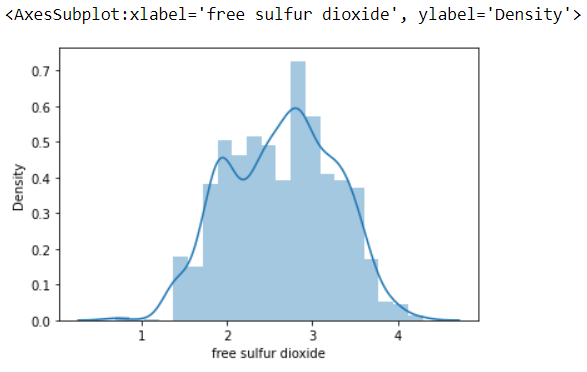


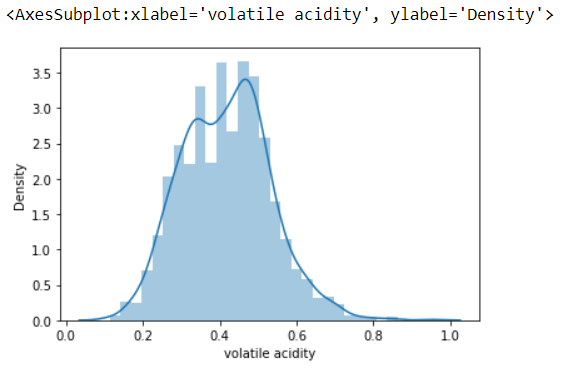
The above image reveals that how that data is easily distributed on features.

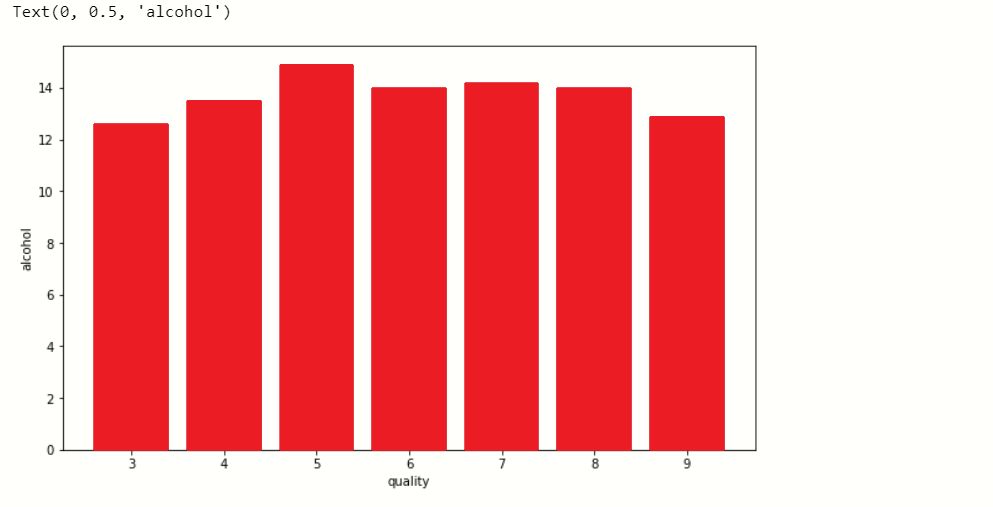
Now, we plot the bar graph in which we check what value of alcohol can able to make changes in quality.





****

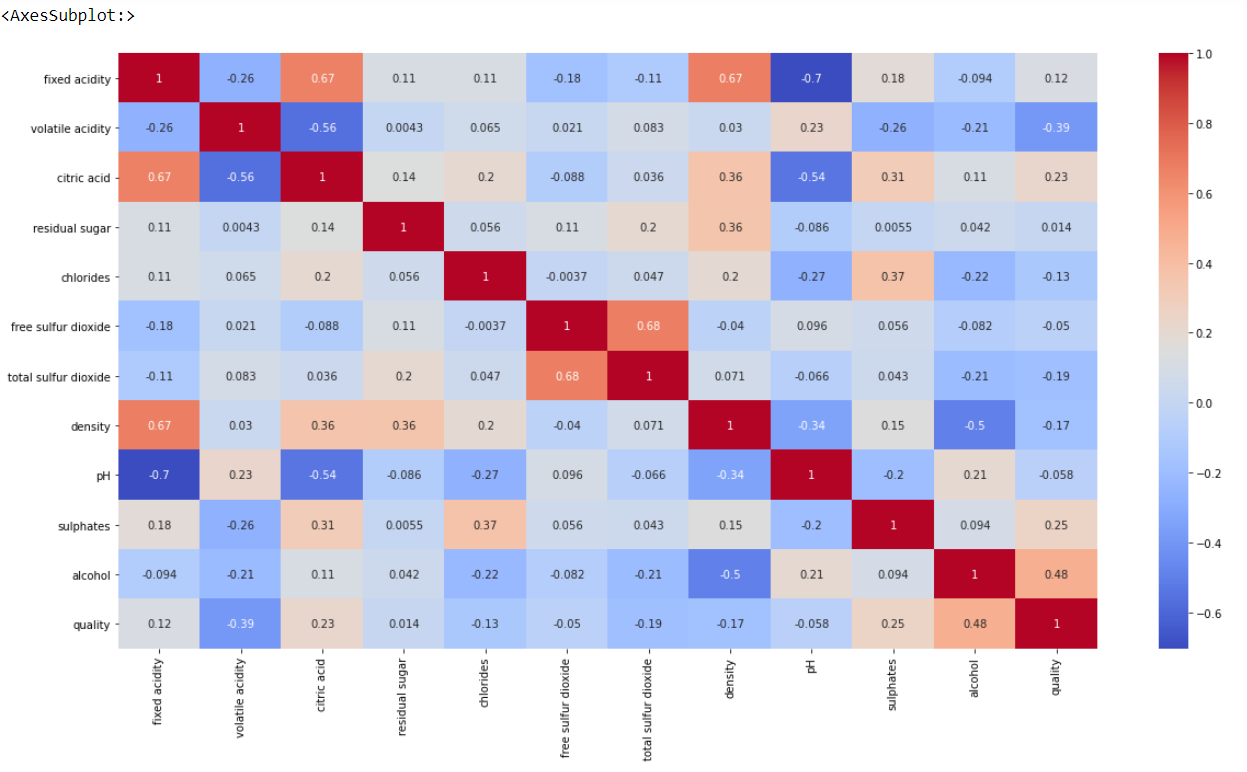
****



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.



Now, we have to find those features that are fully correlated to each other by this we reduce the number of features from the data.

If you think that why we have to discard those correlated, because relationship among them is equal they equally impact on model accuracy so, we delete one of them.

Here we write a python program with that we find those features whose correlation number is high, as you see in the program we set the correlation number greater than 0.7 it means if any feature has a correlation value above 0.7 then it was considered as a fully correlated feature, at last, we find the feature total sulfur dioxide which satisfy the condition.

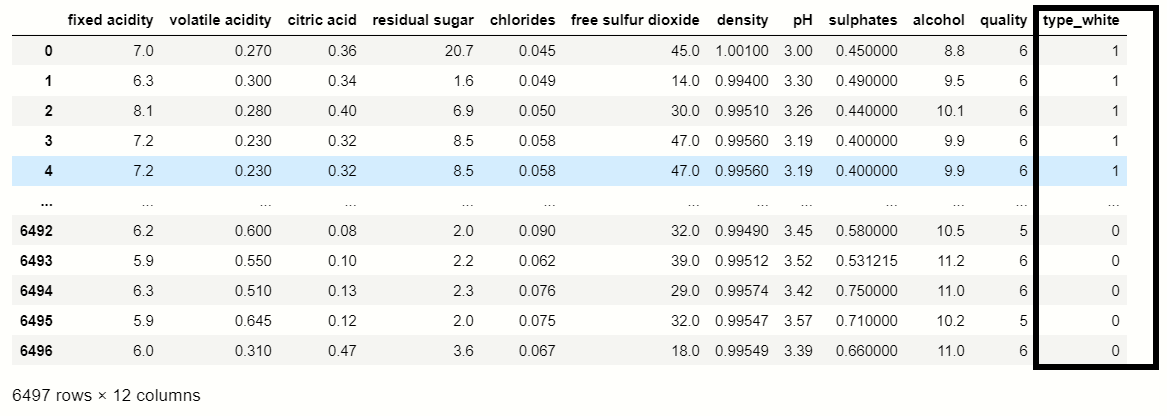
**Pre-Processing Pipeline.**

Handle null values

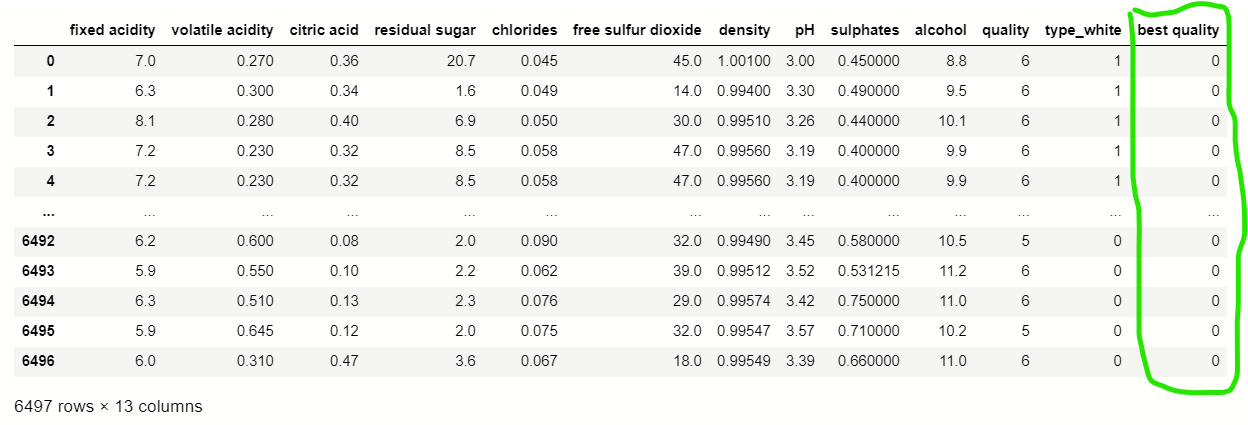
In the dataset, there is so much notice data present, which will affect the accuracy of our ML model. In machine learning, there are many ways to handle null or missing values. Now, we will use them to handle our unorganized data.

We see that there are not many null values are present in our data so we simply fill them with the help of the **fillna()** function.

with this, we handle only numerical variables value because, we fill **mean()**and mean value is not for categorical variables, so for categorical variables:-



You were able to see that the **get\_dummies()** function which is used for handling categorical columns, in this dataset **‘Type’**feature contains two types **Red**and **White,**where Red consider as **0** and white considers as **1.**



**Building Machine Learning Models.:**

We do normalization on numerical data because our data is unbalanced it means the difference between the variable values is high so we convert them into 1 and 0.

 Applying Model

This is the last step where we apply any suitable model which will give more accuracy, here we will use ***Classifier*** because it was the only ML model that gives the 88% accuracy which was considered as the best accuracy.

**Classifier:-**

Training accuracy : 0.7572977481234362

Testing accuracy : 0.7225

precision recall f1-score support

0 0.75 0.69 0.72 204

1 0.70 0.76 0.73 196

accuracy 0.72 400

macro avg 0.72 0.72 0.72 400

weighted avg 0.72 0.72 0.72 400

[[140 64]

[ 47 149]]

Now, we are at the end of our article, we can differentiate the predicted values and actual value.

 Concluding Remarks.

Machine learning skills to solved the issue and The main goal of our project is to solve the problem and predict the wine quality.

Thank you.

Manisha Kumari