ANALYZING WINE QUALITY PREDICTION USING MACHINE LEARNING TECHNIQUES

**DILIP KUMAR**

**DATA TRAINED:B.NO-1829**

**DEPARTMENT OF B.E(ENTC)**

**ACEM,UNIVERSITY OF PUNE,INDIA**

**EMAIL-ID:DK4U90@GMAIL.COM**

**Abstract:**

Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality.

Classification models used here are:

1. Random Forest
2. Gradient Descent
3. SVC
4. Decision Tree
5. K Nearest Neighbor

**Contents:**

Business Understanding …………………………………………………………….1

Data Understanding …………………………………………………………………2

Data Preparation …………………………………………………………………….3

Importing modules…………………………………………………………3.1

Visualization………………………………………………………………..3.2

Handle null values …………………………………………………………3.3

Data Cleaning………………………………………………………………3.4

Data Exploration and Transformation……………………………………...3.5

Split dataset…………………………………………………………………3.6

Feature Importance………………………………………………………….3.7

Normalization …………………………………………………………….3.8

Modeling ………………………………………………………………………….….4

Evaluation ………………………………………………………………………...….5

Save model ……………………………………………………………………………6

End Notes ……………………………………………………………………………7

1. **Business Understanding**

The wine industry shows a recent exponential growth as social drinking is on the rise. Nowadays, industry players are using product quality certifications to promote their products. This is a time-consuming process and requires the assessment given by human experts, which makes this process very expensive. Also, the price of red wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Another vital factor in red wine certification and quality assessment is physicochemical tests, which are laboratory-based and consider factors like acidity, pH level, sugar, and other chemical properties. The red wine market would be of interest if the human quality of tasting can be related to wine's chemical properties so that certification and quality assessment and assurance processes are more controlled.

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.

This project aims to determine which features are the best quality wine indicators and generate insights into each of these factors to our model's wine quality.

To the ML model, we first need to have data for the wine quality dataset. This dataset was picked up from the DT.

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:

1. **Data Understanding**

The Dataset contains a total of 12 variables, which were recorded for 1,599 observations. This data will allow us to create different regression models to determine how different independent variables help predict our dependent variable, quality. Knowing how each variable will impact the red wine quality will help producers, distributors, and businesses in the red wine industry better assess their production, distribution, and pricing strategy.

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.

**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**

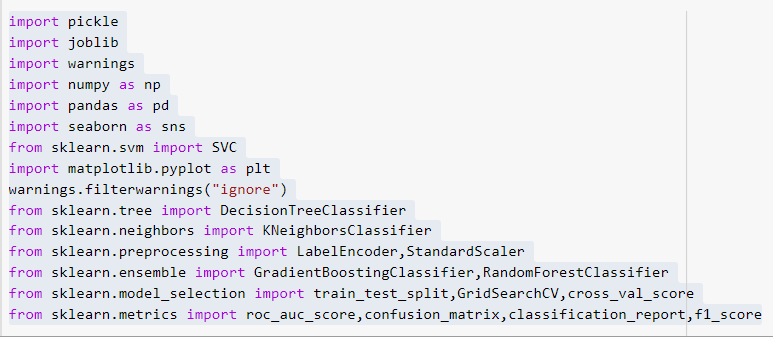
**pH :**   In wine pH is used for checking acidity density

**sulphates :**    Added sulfites preserve freshness and protect wine from oxidation, and bacteria.

**alcohol :**   Percent of alcohol present in wine.

1. **Data Preparation:**

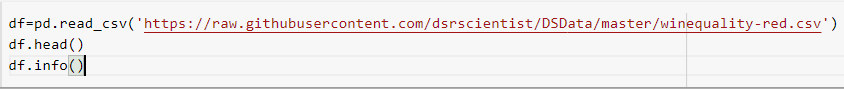
**3.1 Importing modules:**



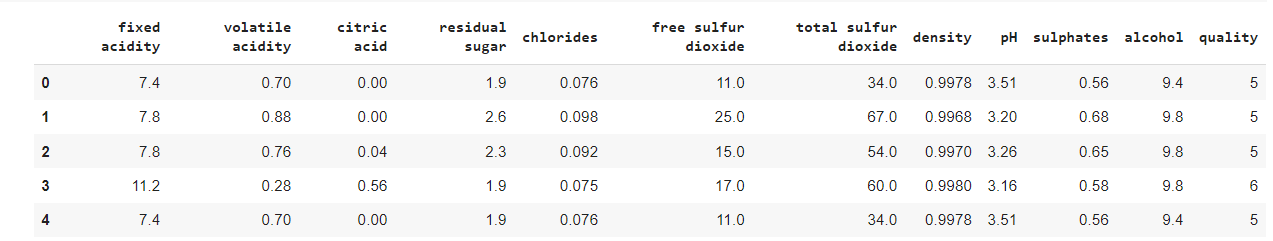
Let’s we take brief about these libraries, pandas are used for data analysis NumPy is for n-dimensional array seaborn and matplotlib both have similar functionalities which are used for visualization.

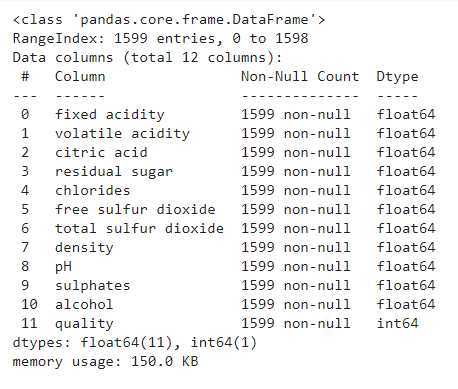
**3.2 Visualization**

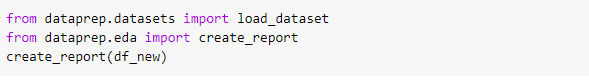
We know that the “image speaks everything” here the visualization came into the work, we use visualization for explaining the data. In other words, we can say that it is a graphic representation of data that is used to find useful information.



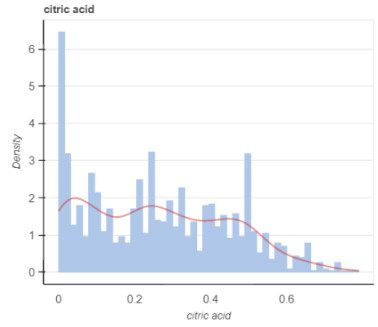
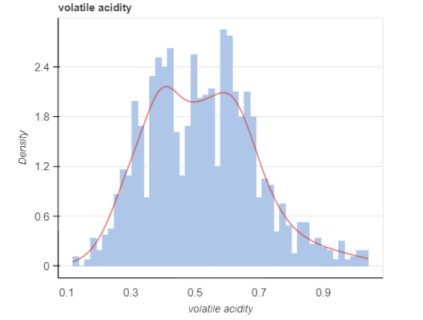
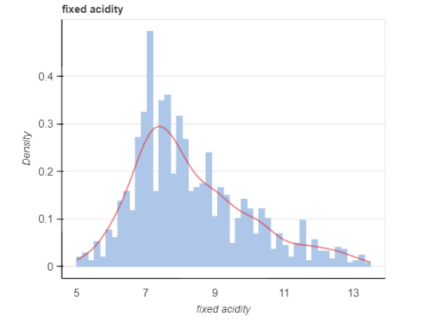
Output:

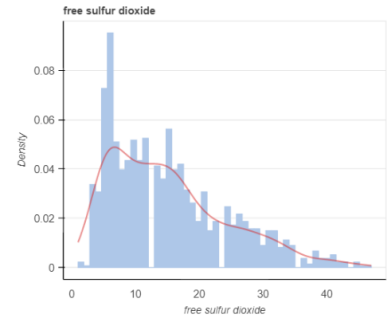
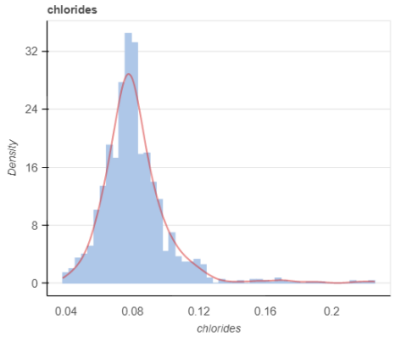
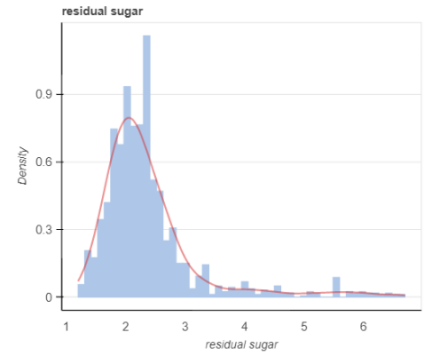


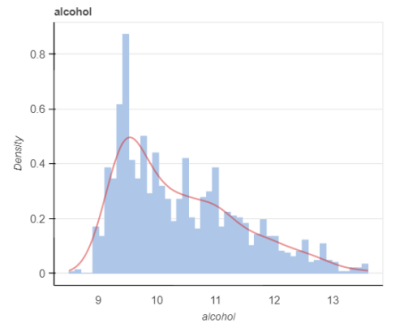
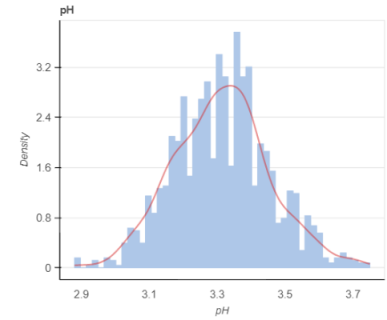
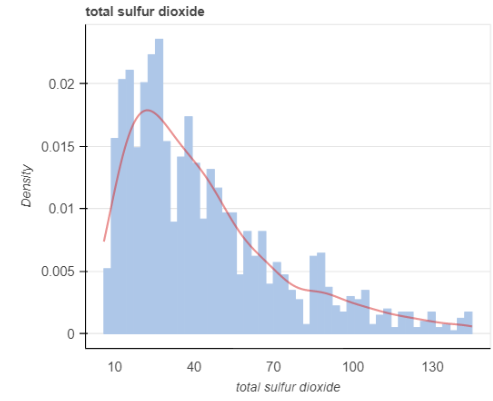




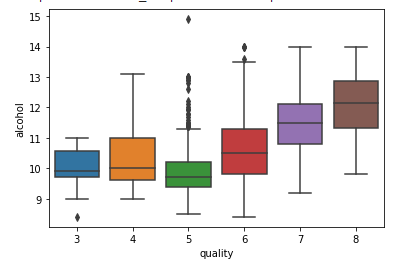
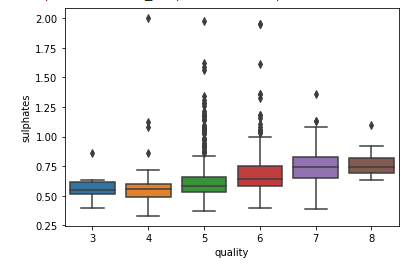
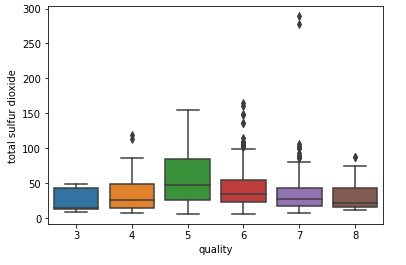
Next, for independent numerical variables, the first step to further analyze the relationship with our dependent variable was to create density plots visualizing the spread of the data

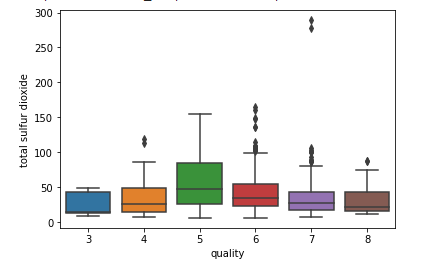
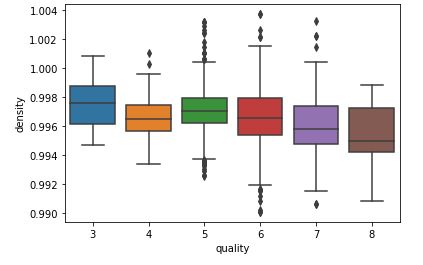
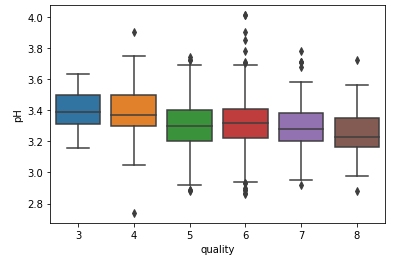


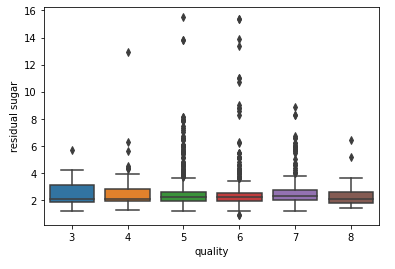
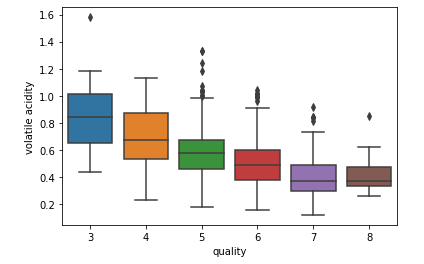
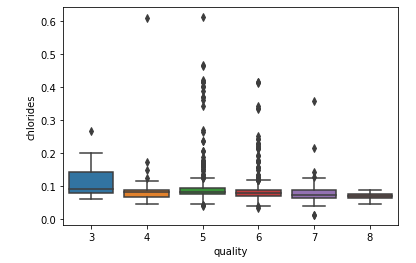


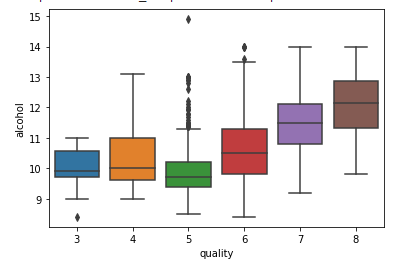
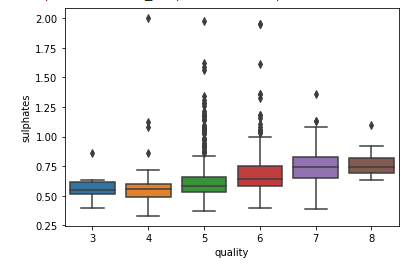
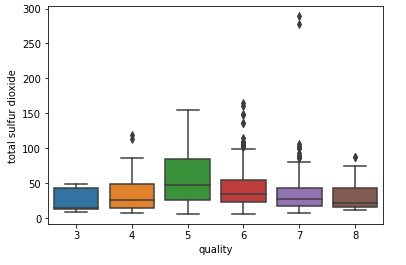


It can be seen that most red wines' pH levels are always between 3-4 and chlorides – the amount of salt is most prevalent at level 0.1. After analyzing the density plots, I plotted the interaction between our numeric variables of interest and our dependent variable of quality





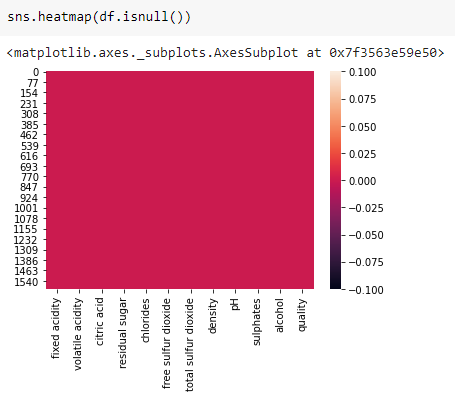
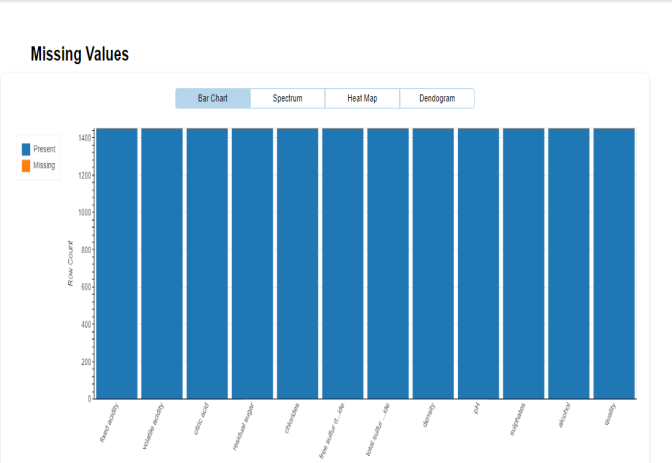




### **3.3 Handle null values:**

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.

There is no any NULL values are present in dataset.

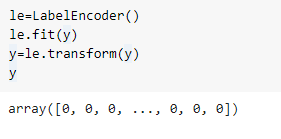
**3.4 Data Cleaning**

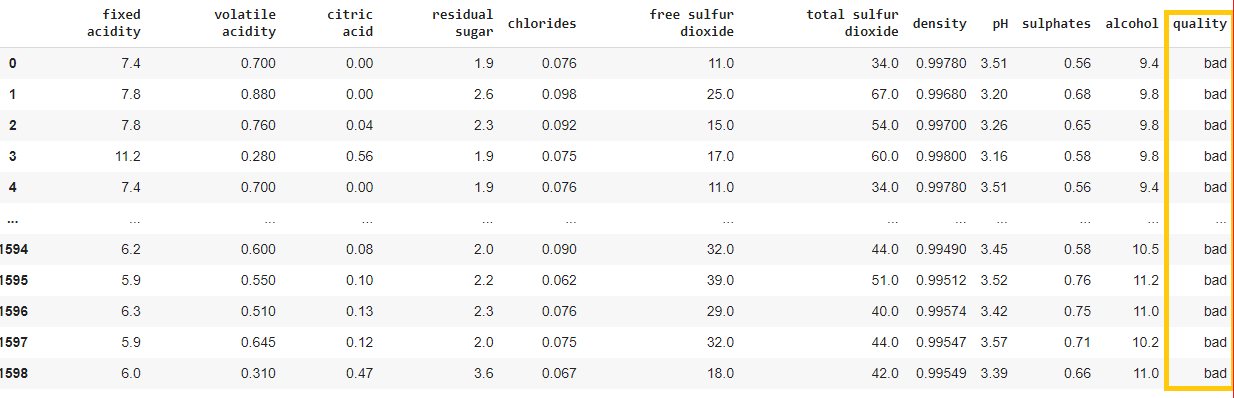
For the next step, we have to check what technical information contained in the data.

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.

**Categorical Feature:**

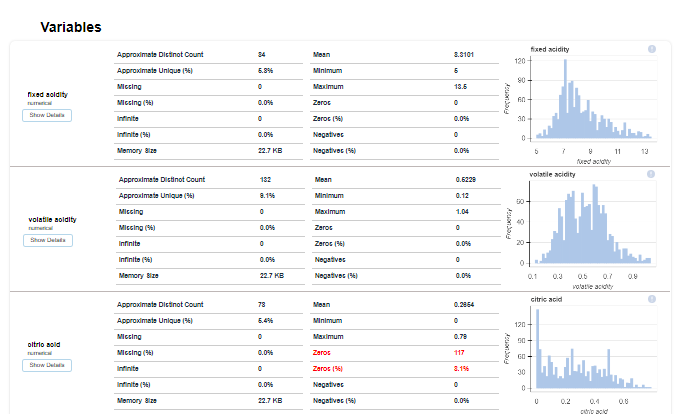
You were able to see that the LabelEncoder() function which is used for handling categorical columns, in this dataset ‘quilityfeature contains two types good and Bad, where good consider as 0 and bad considers as 1.

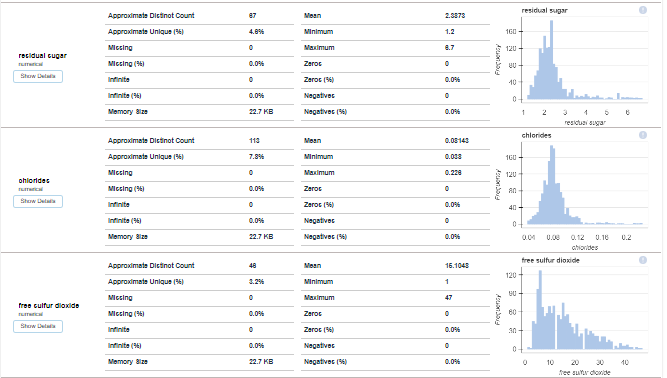


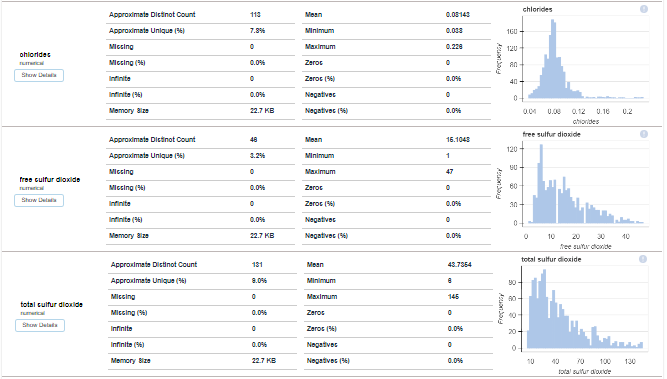


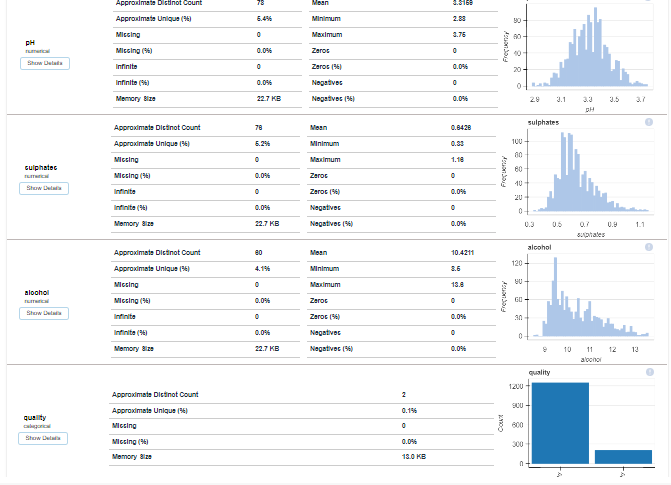
### 3.5 Data Exploration and Transformation:

Data Exploration and Transformation To see which variables are likely to affect the quality of wine the most, I ran a correlation analysis of our independent variables against our dependent variable, quality. This analysis ended up with a list of variables of interest that had the highest correlations with quality.

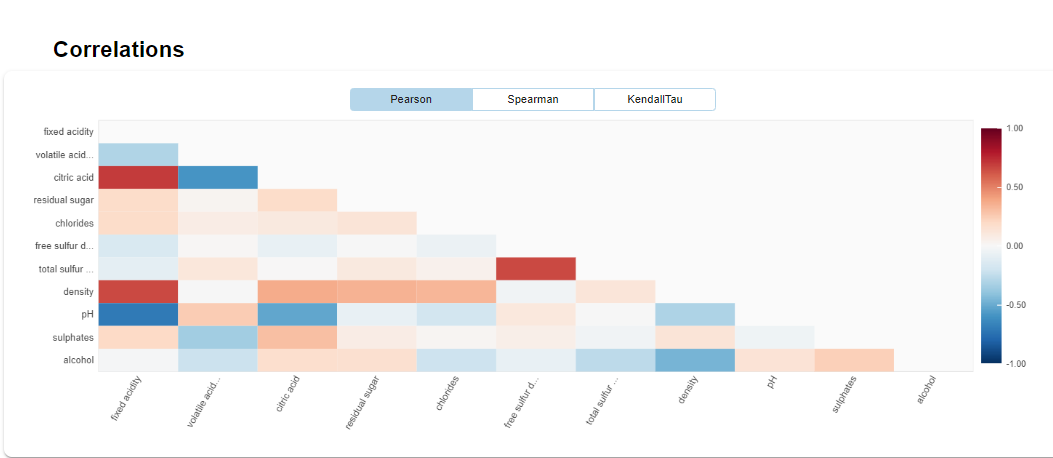


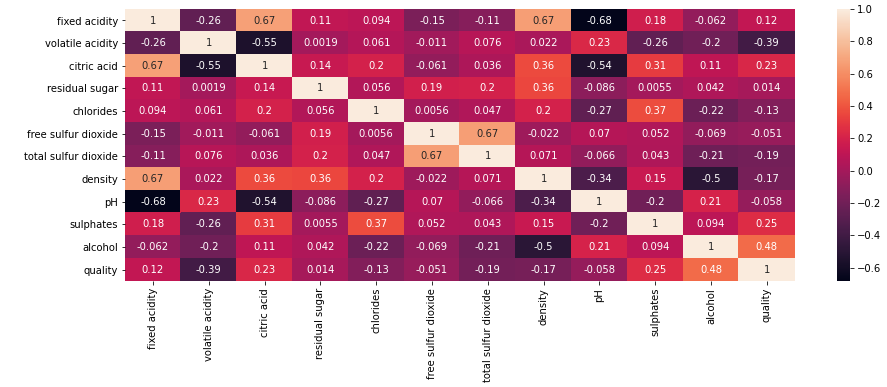






The above image reveals that how that data is easily distributed on features.





A heat map is a two dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using matplotlib tools as per the creators requirement.

In order of highest correlation, these variables are:

1. **Alcohol**: the amount of alcohol in wine
2. **Volatile acidity**: are high acetic acid in wine which leads to an unpleasant vinegar taste
3. **Sulphates**: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
4. **Citric Acid**: acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)
5. **Total Sulfur Dioxide**: is the amount of free + bound forms of SO2

6. **Density**: sweeter wines have a higher density

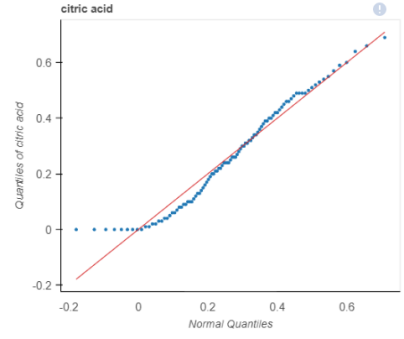
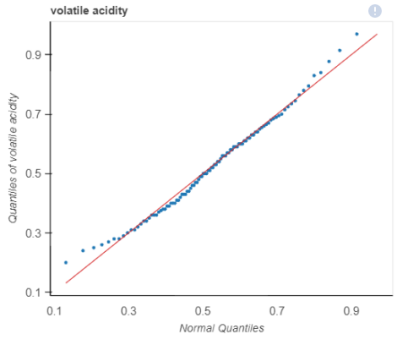
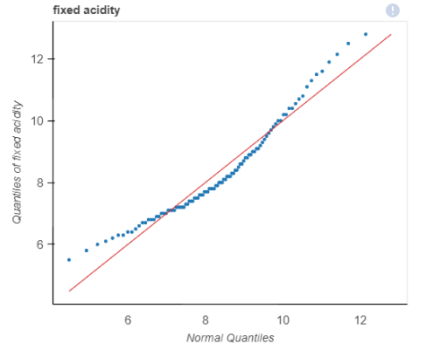
7. **Chlorides**: the amount of salt in the wine

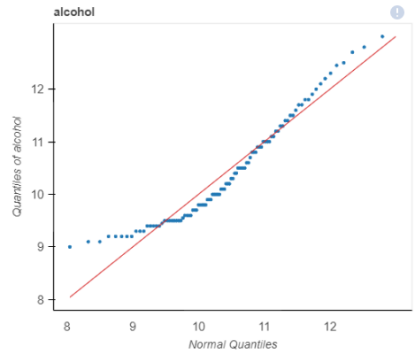
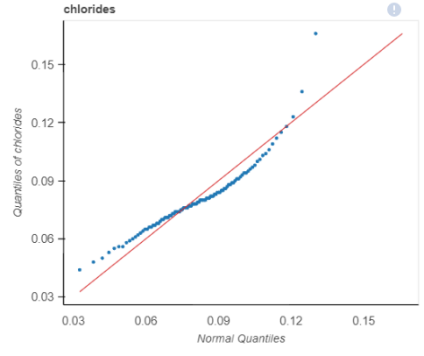
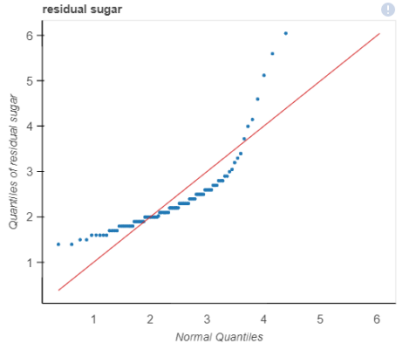
8. **Fixed acidity**: are non-volatile acids that do not evaporate readily

9. **pH**: the level of acidity

10. **Free Sulfur Dioxide**: it prevents microbial growth and the oxidation of wine

11. **Residual sugar**: is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between - sweetness and sourness (wines > 45g/ltrs are sweet)





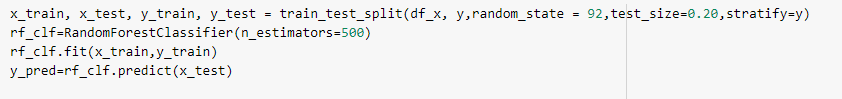
Three different patterns can be observed. First, there are positive relationships between quality and critic.acid, alcohol, and sulphates. Even though wines with a higher level of alcohol may make them less popular, they should be highly rated in quality. Second, there are negative relationships between quality and volatile.acidity, density, and pH. It is reasonable that less sweet wines and a lower level of acidity are favored in quality testings.Last, these independent variables show no significant relationship with quality: residual.sugar, chlorides, and total.sulfur.dioxide.

Interestingly, for wines with an alcohol percentage level below 14, as the level of citric acid increases, there is a rise in red wines' quality. The only exception was at alcohol 14%, where the citric acid level drops as the wine's quality increases. Finally, an interaction analysis using chlorides in relationships with alcohol and quality shows that the wines' quality decreases when chloride level decreases at the alcohol before 12%. However, the quality of red wine increases as the chloride level increases at the alcohol level from 12%

Last, we considered if the collinearity problem existed in our data. As a result of correlation analysis and VIF verification, we discovered some variables with slightly high correlations.

**3.6 Split dataset:**

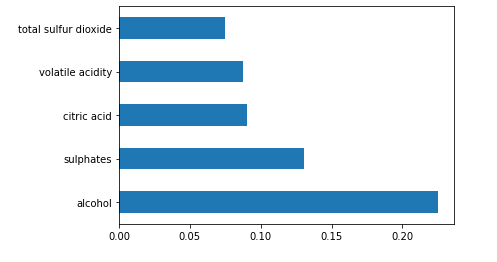
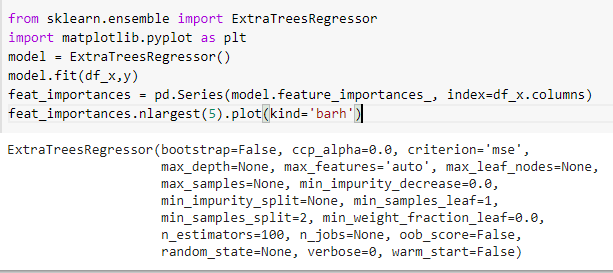
Now we perform a split operation on our dataset.



**3.7 Feature Importance:**

Another great quality of random forest is that they make it very easy to measure the relative importance of each feature. Sklearn measure a features importance by looking at how much the treee nodes, that use that feature, reduce impurity on average (across all trees in the forest). It computes this score automaticall for each feature after training and scales the results so that the sum of all importances is equal to 1. We will acces this below:

Top Five Feature importance Graphysically representations

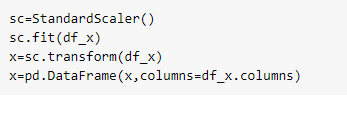
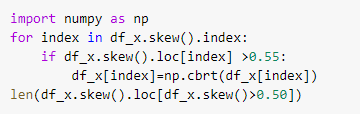


As per aboveb performance metrics, classification report and prediction confusion matrix of the wine data using decision tree classifier. Which classifies the quality of the wine based on the properties and give output as high, low and medium.

**3.8 Normalization:**

My first step was to clean and prepare the data for analysis. I went through different steps of data cleaning. First, I checked the data types focusing on numerical and categorical to simplify the correlation's computation and visualization. Second, I tried to identify any missing values existing in our data set. Last, I researched each column/feature's statistical summary to detect any problem like outliers and abnormal distributions.

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.

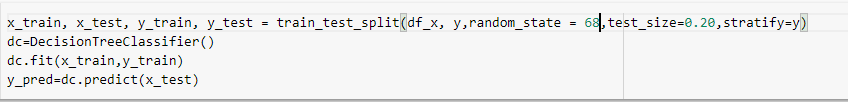


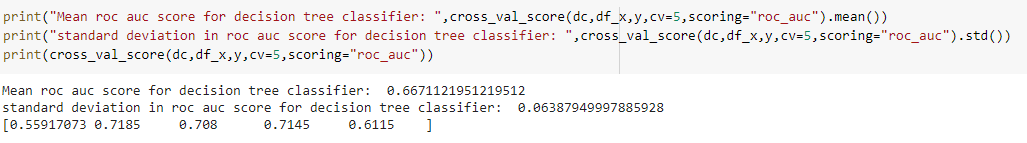
Last, we considered if the collinearity problem existed in our data. As a result of correlation analysis and VIF verification, we discovered some variables with slightly high correlations.

1. **Applying Model**

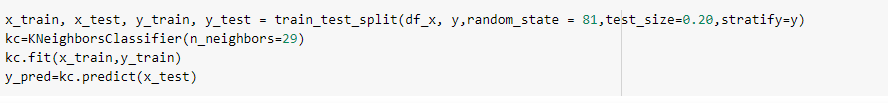
Now we will train several Machine Learning models and compare their results. Note that because the dataset does not provide labels for their testing-set, we need to use the predictions on the training set to compare the algorithms with each other. Later on, we will use cross validation.

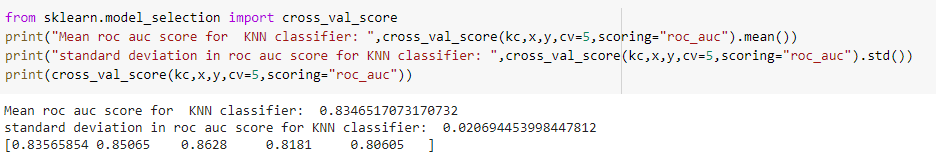
**Decision Tree**



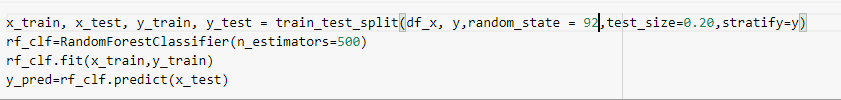


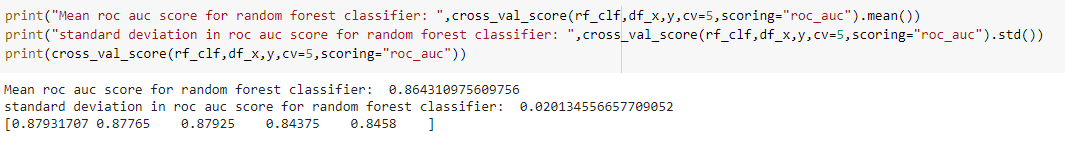
**K Nearest Neighbor**



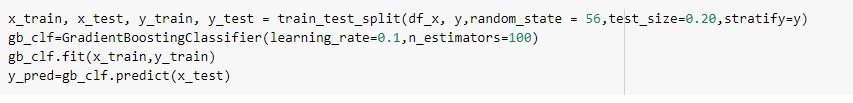


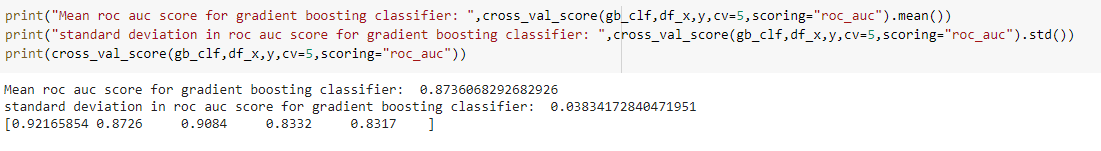
**Random Forest:**



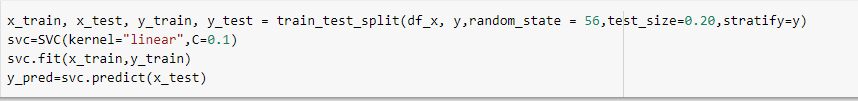


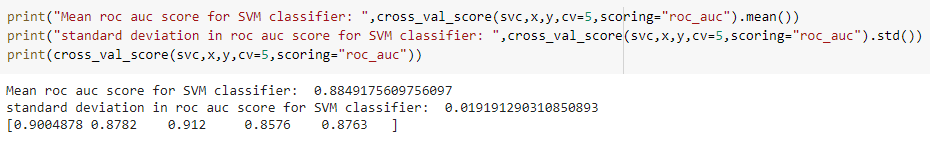
**Gradient Descent (GD):**





**Linear Support Vector Machine:**

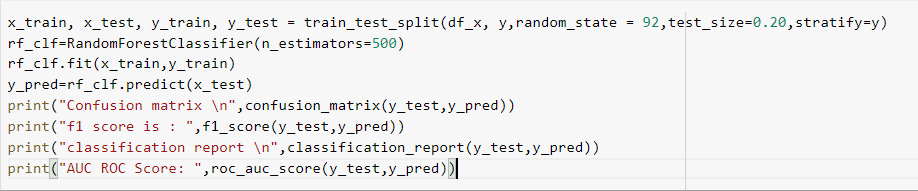




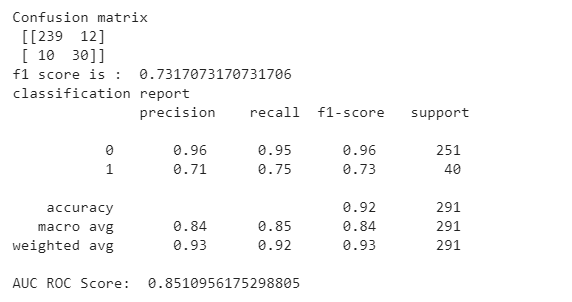
**Final Model Selection based on accuracy and ROC AUC score is RandomForestClassifier**

Based on above scores we can choose random forest as there we are achieving highest auc\_roc score with random state 92

This is the last step where we apply any suitable model which will give more accuracy, here we will use RandomForestClassifier because it was the only ML model that gives the 86% accuracy which was considered as the best accuracy.



Output:



gives the performance metrics, classification report and prediction confusion matrix of the wine data using random forest classifier which is also used to classify the quality of the wine.

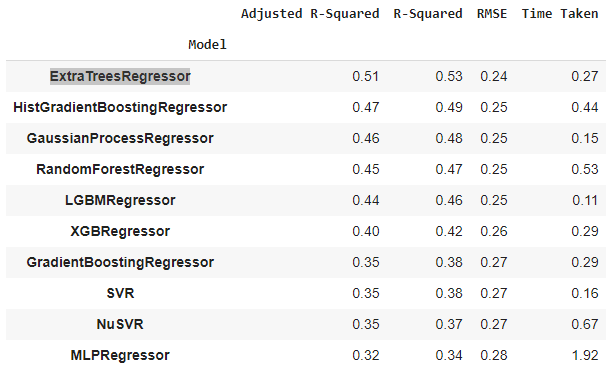
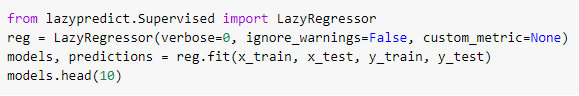
### **Applying Lazy Prediction**

One of the problems of the model-building exercise is ‘How to decide which machine learning algorithm to apply ?’

This is where Lazy Prediction comes into the picture. Lazy Prediction is a machine learning library available in python that can quickly provide us with performances of multiple standard classifications or regression models on multiple performance matrices.

Let’s see how it works…

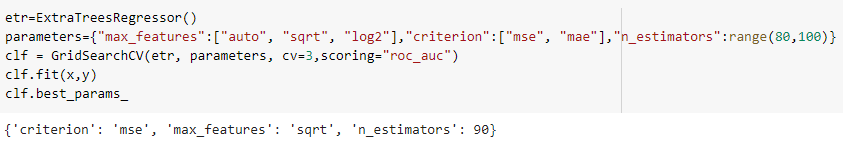
Since we are working on a Regression task we will use Regressor models.



As we can see LazyPredict gives us results of multiple models on multiple performance matrices. In the above figure, we have shown the top ten models.

Here ‘ExtraTreesRegressor’ outperform other models significantly. It does take a high amount of training time with respect to other models. At this step we can choose priority either we want ‘time’ or ‘performance’.

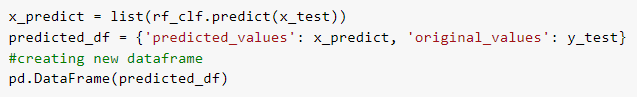
We have decided to choose ‘performance’ over training time. So we will train ‘ExtraTreesRegressor’ and visualize the final results.



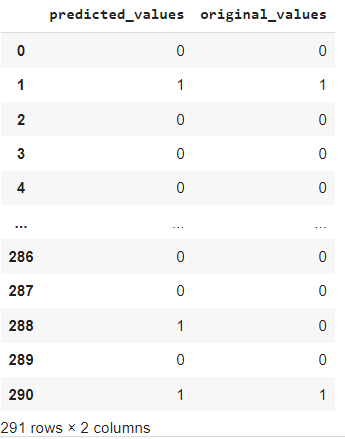


1. **Evaluation:**

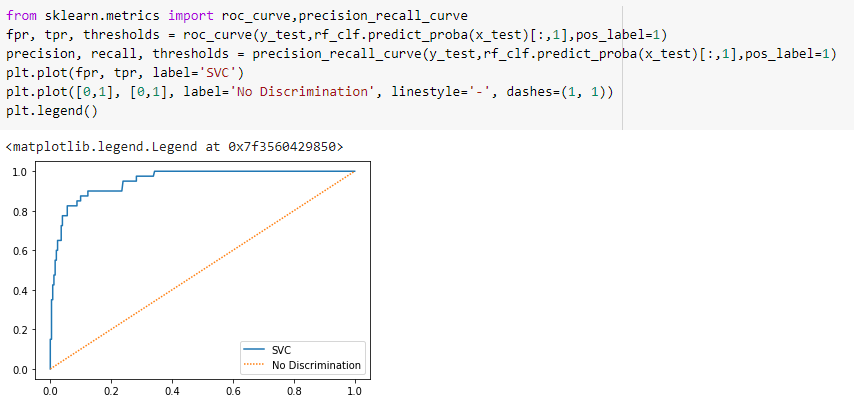
Now, we are at the end of our article, we can differentiate the predicted values and actual value.



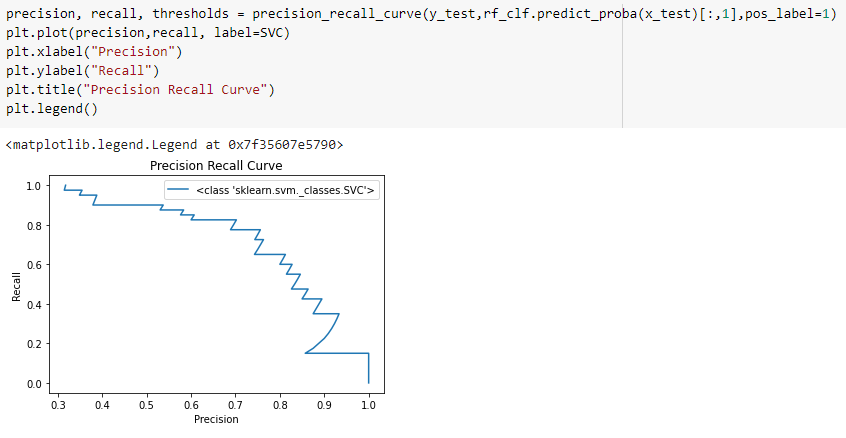
Output:



**ROC Curve**

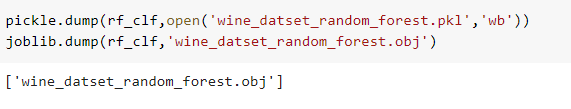


**Precision Recall Curve**



### **Saving Model**

At last, we save our machine learning model:



### **End Notes:**

This is one of the interesting articles that I have written because it was on today’s current top technology machine learning, but I was used basic language to explain this article so, you can’t get difficulty on understanding.

The specific objective of this study is to analyse how physicochemical properties like alcohol percentage, chlorides, sulphates contents etc., varies the quality of wine. This study analyse the wine types and quality with the various physicochemical variables. Out of all attributes, the statically significant attribute that influence the quality of wine is an essential finding. The model that highlights the significant attribute . This result helpful in production and in quality prediction by studying those attributes. Analyse the wine quality by different machine learning algorithms such as decision tree, random forest and extreme gradient boosting. The results obtained are more accurate than previous techniques.