***WINE QUALITY PREDICTION USING MACHINE LEARNING***

Graphical user interface, text, application

Description automatically generated

Yes, we are going to discuss about WINE and its quality and how we can find the quality of wine?

Generally, Wine is made up of crushed grapes and some physicochemical substance’s such as citric acid, volatile acid, sulfur di oxide, sulphates, alcohol and PH and so on.

Fake wines which are not a quality one where they do not use grapes at all and instead of grapes, they will add sugar and other substance like ethanol to enhance the taste to the original one.

So, in this case, the quality of wine changes for sure and how we can differentiate it?

“*Experts confirmed that wine quality can be identified by its smell, color and here our topic is to how we can find the wine quality”.*

Using ML techniques, we can find the quality of wine. Let’s see how we can predict the quality of wine,

**DATA DESCRIPTION:**

We have **FEATURES** and **TARGETS** which is nothing but the Input variables and Output variables. We need Features to predict the Target.

Here our target variable is Quality, and we need to predict the quality of Given Red/White Wine.

Below are the Input Variables i.e., Features,

*1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 – alcohol*

**Importing Required Libraries:**

Here We are importing the required libraries for this prediction,

Graphical user interface, text, application

Description automatically generated

* Pandas is open-source library tool which provides high performance data analysis tool by its powerful data structures.
* It helps to shorten the procedure of handling the data with extensive set of features.
* NumPy is most used package for scientific computing for multi-dimensional array of objects.
* Data will be understandable more in plots or graphs or charts than in raw form. Agree right? So, we are bringing the below packages for that,
* Matplotlib and Seaborn are the 2 data visualization libraries which provides a stunning plot for visualizing the data for better understanding.
* Apart from this we are importing Metrics which is necessary for this project to calculate the score. As it is a classification problem, I am using Classification report, accuracy score / f1 score, Confusion Matrix, roc curve and roc auc score.
* In terms of preprocessing, we are going to use Standard Scaler to scale our data to bring up the range of values between 0 to 1.
* Train test split for splitting the train data and test data for building the model and predicting the results.

**Understanding the Data:**

After loading the data, we will get the basic understanding of our data like whether our data features are numerical values or categorical values and what kind of values it has and so on…

Table

Description automatically generated

As we can see that our data has 1600 rows and 12 columns including target variable which is “Quality “and all our data is numerical values.

The data do not have any missing values and the below function proves that as well,

Text

Description automatically generated

**Data-Visualization:**

We will see how the data has been distributed in the dataset to have some knowledge on Outliers or skewness if any present in dataset.

We used distplot for showing the data on how it is distributed on the dataset.

This is example of the Univariate distribution of data which is data is distributed against the density.

Shape, polygon

Description automatically generated

Most of the columns in our dataset has skewness/ outliers present in our dataset and we can see that it is skewed positively except density, PH, Quality column which has skewness in both positive and negative.

Applying Z score transformation method to treat outliers and we have lost ~9 % data which is acceptable as data loss is acceptable <10%.

Text

Description automatically generated

**Correlation Checking:**

Heat Map is a rectangular data, and it shows a two-dimensional view with color – encoded matrix.

Heat Map plot is used mostly for checking correlation between features vs features and features vs target.

If there is any correlation exists, we can drop one of the features based on the importance which has on the dataset for prediction the target variable.

Generally, the value which is having more than 0.5 is considered as correlation exist and if its more than 0.7 – 0.8, then it’s a strong correlation and we can drop one of features.

We can use scatter plot for the 2 columns and check further on correlation exists or not.

Chart, treemap chart

Description automatically generated

As we can see features like PH, citric acid, fixed acidity looks like they have correlated to each other, but I am not dropping here as it would be the important features in our dataset to test the quality of wine.

Most interesting thing here is that variable – Quality has values from scale 1 – scale 10. So, I need to check whether the quality of wine is Good or Bad. Our Prediction Model is **Binary Classification Method**, So I need to Convert as below in order to do prediction.



The wine quality which is having values >=7 is considered as 1 which is GOOD else 0 which is BAD. Now its easy to predict the quality of wine.

So, our Exploratory Data Analysis part is done!!! **😊**

Let’s check the target variable through plot to check whether target classes are balanced or not. If it is not balanced, we should make it balanced before building a model for prediction.

Chart, bar chart

Description automatically generated

As expected, our classes are im-balanced ☹ and we need to balance these classes and it has 2 ways,

1. Increasing the minority class to majority class  OVER – SAMPLING (*most preferred*)
2. Decreasing the majority class to minority class  UNDER – SAMPLING (*Not preferred as it will delete the data from majority class and make it balanced to minority as it results in loss of data*).

So, I am Using Over sampling SMOTE technique to balance the classes. please ***install imblearn*** in order to apply the re-sampling technique.

Text, letter

Description automatically generated

Now our classes are balanced, and we can proceed to the next step which is Scaling the data.

Why we need to do scaling?

* We will have different range of data - mg, mm etc. So, it will cause the bias. So, in order to equalize the range of values, we are doing scaling before training our model.
* So, the scaled data will have values ranges from 0 to 1 and we should do Scaling only for features (Input Variables) not for targets.



**Train-Test-Split:**

This is important step where we are splitting our data for training the model and using test data, we will predict the model score.

From the whole dataset which is generally 100%, I am splitting 25% to test data and remaining 75% will be allocated to train data. We can allocate max 30% to test data.



Here comes the funny and most important part, which is Model Building,

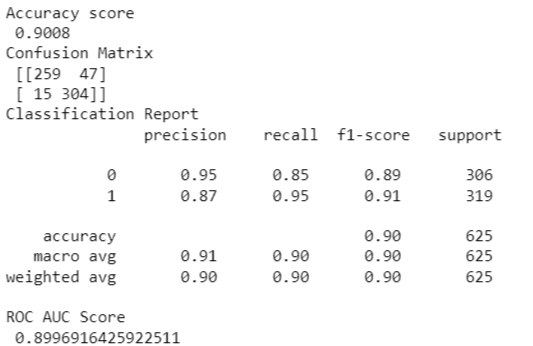
**Model Building:**

Every time it is good to build multiple models for the dataset as some models will work well for the given dataset and some will give less accuracy.

Train the data with multiple models and choose the best model from that.

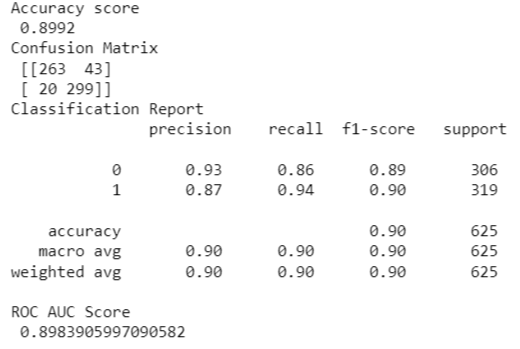
1. **Support Vector Classifier – F1 Score -86%**

Goal of SVC is to divide the datasets into classes to find the maximum marginal hyperplane. Data points which are close to the hyper plane are Support vectors.



1. **Decision Tree Classifier – F1 Score -85%**

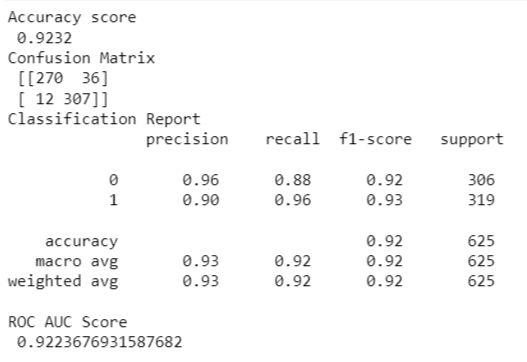
It is one of the most powerful tools for classification Prediction and it is a flow chart like tree structure which has branches, leaf node, root node.



1. **Random Forest Classifier – F1 Score- 92%**

It is one of the Ensemble classification algorithms which operates by constructing multiple decision trees at training time.

It will predict by taking average or mean of the output from various trees.



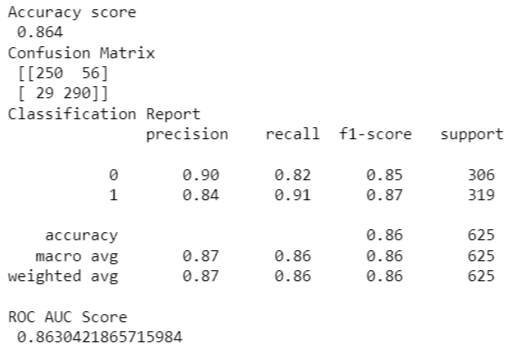
1. **Ada Boost Classifier – F1 Score- 86%**

Ada Boost (or) Adaptive Boosting algorithm is one of the boosting classifier methods used to increase the accuracy of model.

It will combine multiple poorly perform classifiers so that you will get high accuracy classifier.

The advantage is that it is easy to implement, and it is not prone to over fitting.

The dis-advantage is that it is sensitive to noise data and highly affected by outliers and it is slower when compares to XGBoost which is Extreme Gradient Boost.



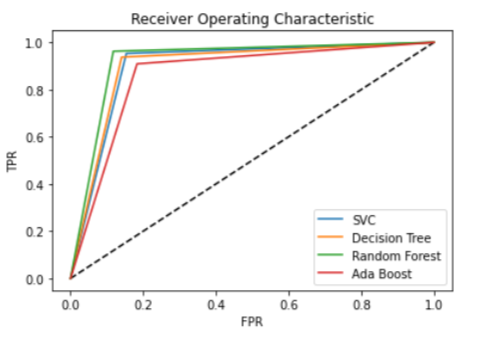
When Our target variables classes are not balanced, we should use **F1 Score** metrics for checking the accuracy of model.

We should **not** check **Accuracy Score** metrics for checking the accuracy of model as it will taken only the majority classes in consideration for calculating the score.

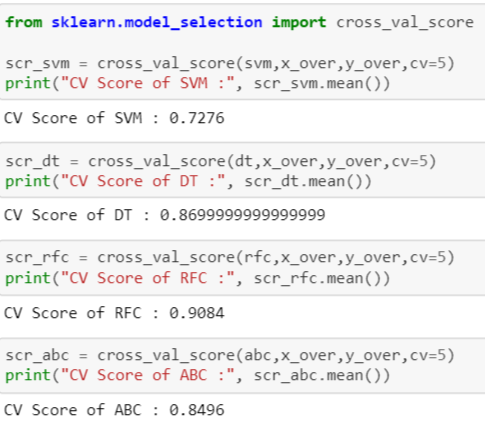
**ROC curve:**

Below are the ROC Curve for all the models and as we used classification method, ROC curve is important, and it will show the Area under curve for the models.

The curve which is closed to top left corner indicates the better performance model which is RANDOM FOREST here.



Use **CROSS- VALIDATION- SCORE** to check whether the accuracy (F1 Score here in this case) of model is over-fit or under- fit.



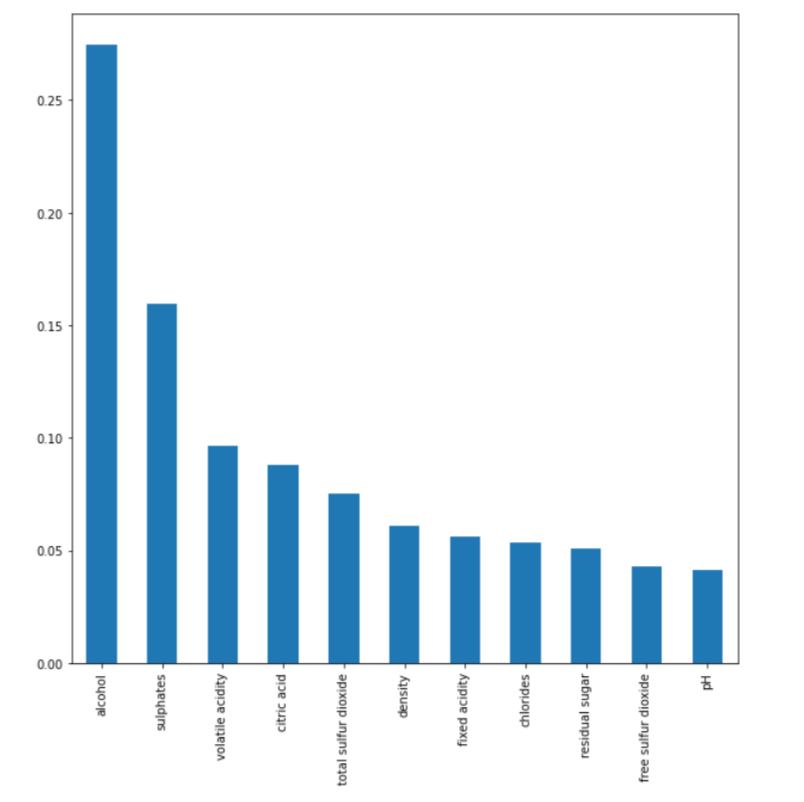
So, we need to choose the best model on checking the difference between cross Val score and model accuracy score. Less Difference makes the best model.

Here **ADA BOOST** is considered as best model as it is having less difference.

Let’s try applying Hyper parameter tuning to improve the accuracy,

**Feature Importance:**

We will look into features importance which is nothing, but the Which features are very important, and it will help us to drop the less important features from the dataset.



Here In this, we can see that Features – Alcohol, Sulphates, Volatile acid, citric acid, Total Sulfur dioxide are having more importance in dataset.

Less Important Features are – pH, Free sulfur dioxide, residual sugar is having least importance in dataset.

Random forest and any boosting algorithms works well for checking the importance of features.

Each Algorithm has different parameters for tuning the accuracy and ada boost has the below accuracy and we have 2 ways to tune the accuracy,

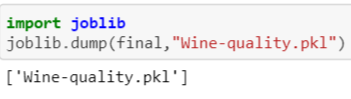
* Grid Search CV
* Randomized Search CV

Randomized Search CV works well in large datasets as it is fast than Grid Search CV. I have used Grid Search CV and **My final score increased to 89%** which is previously 84%.



**Saving the Model:**

One of the most important steps is saving the model. We will save the model as .pkl file as below.



The file will be saved in our system directory and while loading the data at the time of deployment, Developers will load this .pkl file to load the data and predict the result.

Please use the below link to download the dataset,

<https://github.com/dsrscientist/DSData/blob/master/winequality-red.csv>

You can also refer the below GitHub link,

<https://github.com/bramee/Practice-2/blob/main/Wine_Quality_Project.ipynb>