Machine Learning

Wine Quality Predictor

Introduction

About wine:

Wine is a beverage made from fermented grape and other fruit juices with lower amounts of alcohol content. Quality of wine is graded based on the taste of wine and vintage. This process is time taking, costly and not efficient. A wine itself includes different parameters like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol and quality.

Objective:

Our main objective is to predict the wine quality using machine learning.A large dataset is considered and wine quality is modelled to analyse the quality of wine through different parameters like fixed acidity, volatile acidity etc. All these parameters will be analysed through Machine Learning algorithms like random forest classifier algorithm which will helps to rate the wine on scale 1 - 10 or bad - good or 0-1.

Data-Description:

The dataset used is Wine Quality Data set from UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Input variables are *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol*.

Output variable is is quality (score between 0 and 10).

PREVIOUS WORK DONE IN THIS FIELD

Some previous work done in this field is by Apoorva Dave. Here she used this dataset and *predicted the quality of red wine* on a scale of 0–10 given a set of features as inputs. She have solved it as a regression problem using ***Linear Regression.***

Her work can be seen in her github repository.

Implementation (Code)

Code is in the Python Notebook

Experimental Results

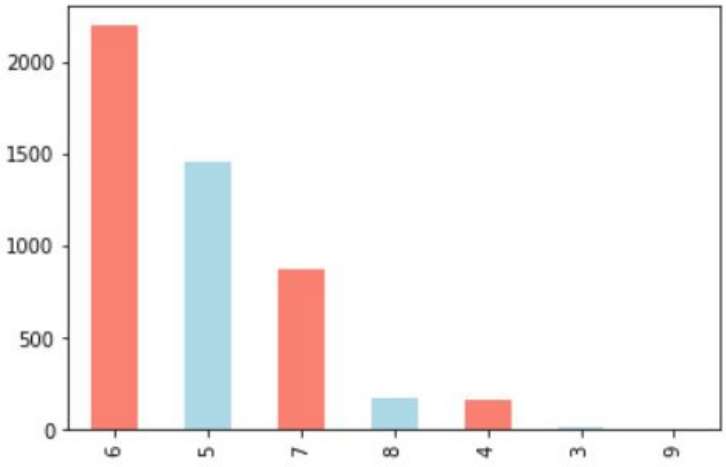
Final results:

The Model Predicts the quality as 0 or 1. 0 Being bad and 1 Being Good.

The Model is also exported out as a “joblib” file, so it can be used to predict in different systems or servers.

Graphs:

1. Value Counts Graph:



Plots Series containing counts of unique values.

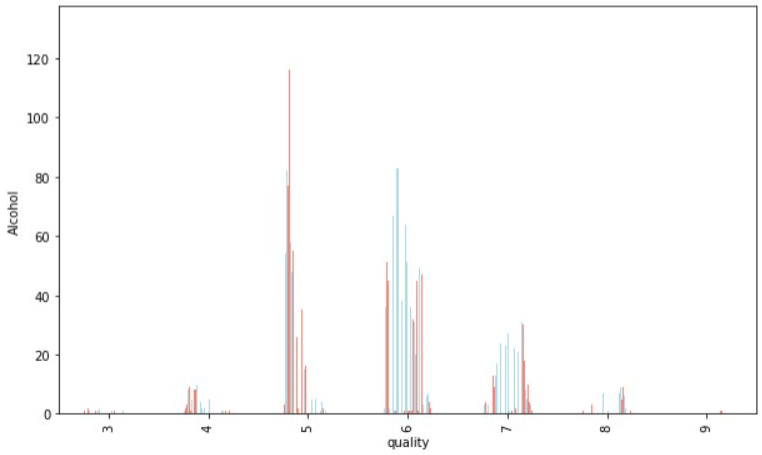
2. Relationship Graph

Relation between Alcohol and Quality of wine

```
]: pd.crosstab(data.quality, data.alcohol).plot(kind="bar",
                                               figsize=(10,6),
                                               color=["salmon", "lightblue"],legend = False);

   plt.ylabel("Alcohol")
```

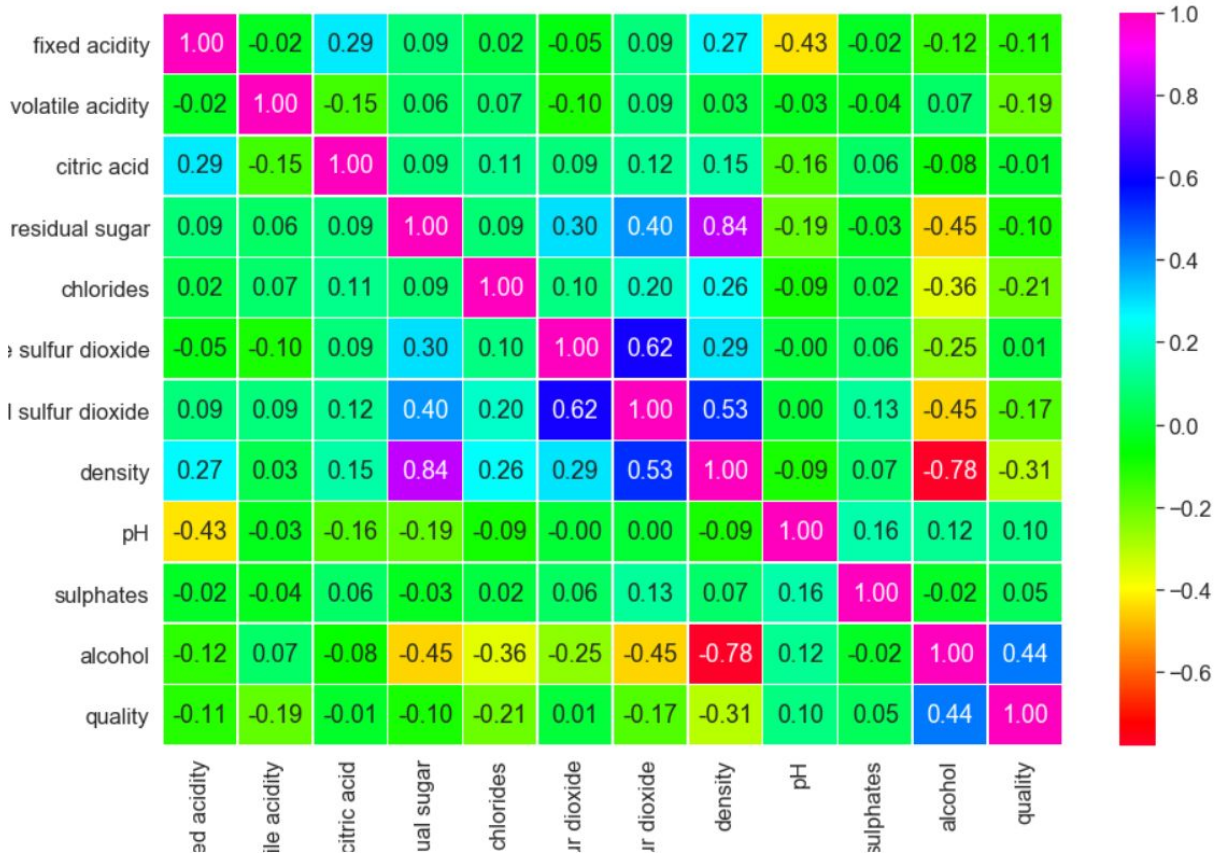
```
]: Text(0, 0.5, 'Alcohol')
```



3. Correlation Matrix Graph;

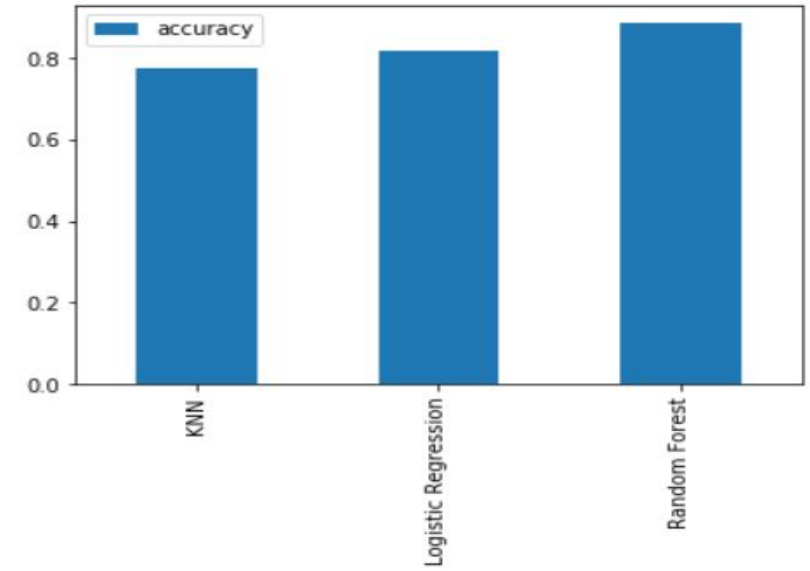
   A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

   The graph is simply plotting the co-relation

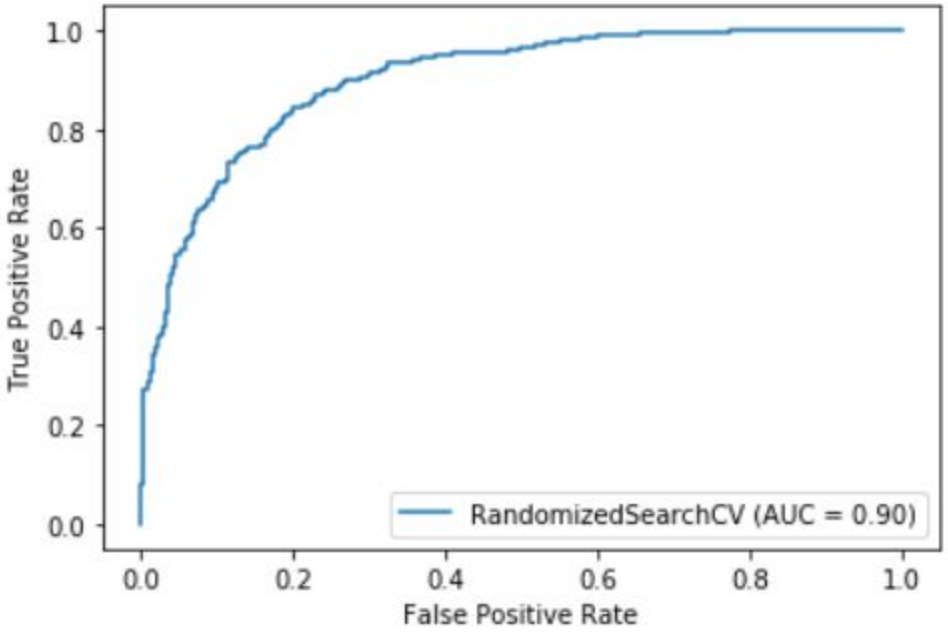


4. Model Accuracy Graph

   As I have tested 3 models to choose from, the following graph shows the accuracy rate of each mode.

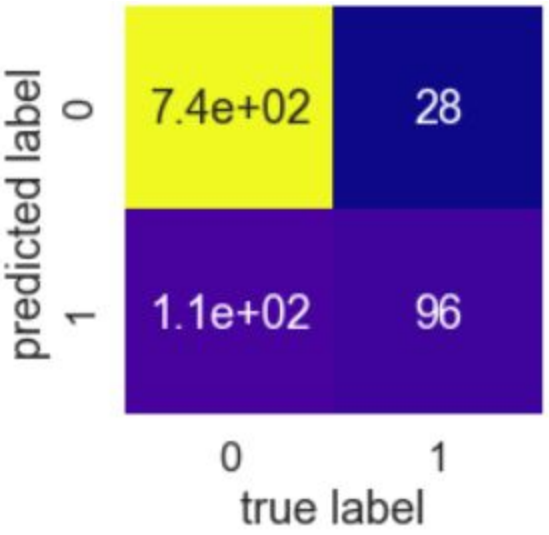


5. ROC AND AUC CURVES



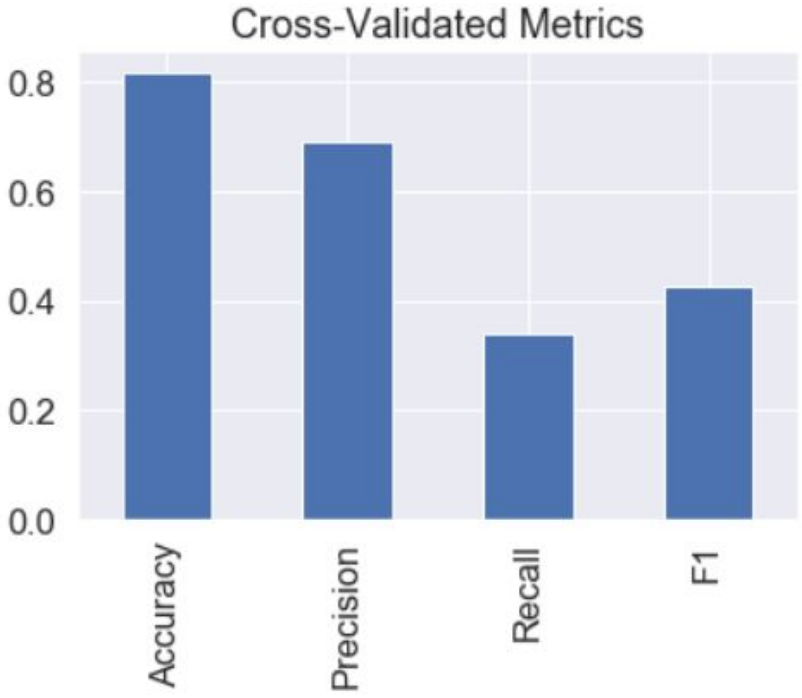
A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate."

6. Confusion Matrix



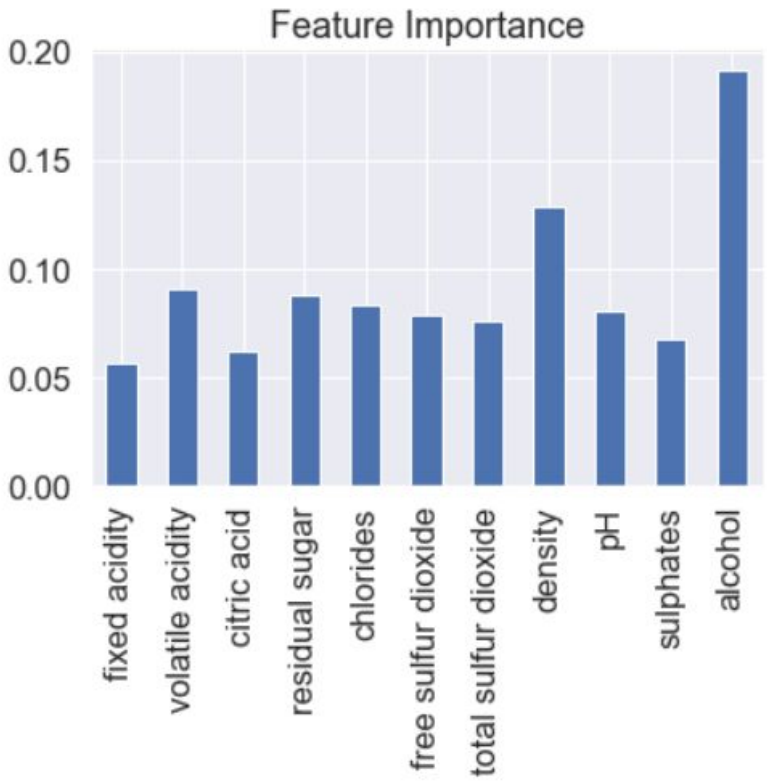
In this the diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

7. Cross Validation graph on different parameters



It just compare different results of the final, hypertuned model

8. Feature Importance



It is a graphical version of which importance of each feature for result

References:

https://scikit-learn.org/stable/index.html

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

https://numpy.org/doc/

https://pandas.pydata.org/docs/user_guide/index.html#user-guide

http://archive.ics.uci.edu/ml/datasets/wine

https://seaborn.pydata.org/tutorial.html

https://vwo.com/blog/heatmap-colors/