Course Title: Data Mining and Machine Learning Lab

Course Code: CSE322

Project Report: Wine Quality Prediction

Submitted To: Md. Zahid Hasan

Associate Professor,CSE

Daffodil International University

Submitted By:

Name: B.M.Samiul Haque Real

ID: 201-15-3057

Section: PC-A

Department of CSE

Wine Quality Prediction

Introduction:

The quality of wine is extremely essential to both consumers and the manufacturing industries. Product quality certification is helping businesses increase their sales. Nowadays, wine is a widely consumed beverage all over the world, and enterprises rely on product quality certification to boost their market worth. Previously, product quality testing was done at the conclusion of the manufacturing process, which is a time-consuming procedure that necessitates a lot of resources, such as the requirement for numerous human specialists to assess product quality, making this process highly expensive. Every person has an opinion about the test, therefore determining the wine's quality based on human specialists is a difficult undertaking. There are various factors that may be used to forecast wine quality, however not all of them are significant for improved prediction.

Problem statement:

In industries, understanding the demands of wine safety testing can be a complex task for the laboratory with numerous analyses and residues to monitor. But, our application’s prediction, provide ideal solutions for the analysis of wine, which will make this whole process efficient and cheaper with less human interaction.

Motivation:

Our main Motivation is to predict the wine quality using machine learning through Python programming language .

• A large dataset is considered and wine quality is modelled to analyses the quality of wine through different parameters like fixed acidity, volatile acidity etc.

• All these parameters will be analyzed through Machine Learning algorithms like random forest classifier algorithm which will helps to rate the wine on scale 1 - 10 or bad - good.

• Output obtained would further be checked for correctness and model will be optimized accordingly.

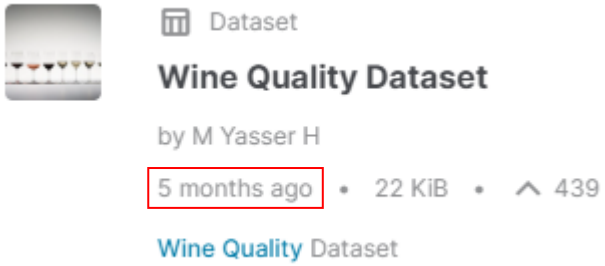
• It can support the wine expert evaluations and ultimately improve the production.

About Dataset:

We employed the algorithms of Logistic Regression, KNN, Decision Tree, and Random Forest. Our dataset has around 1599 records. It has a total of 12 columns. The columns are as follows: (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, experience, pH, sulphates, alcohol and quality). As a null value, we have 0% missing data. Label encoding was one of the things we focused on. It aids in the conversion of a string to a numerical value. We used 30% of the data for the test and 70% of the data for the train.

Acknowledgements:

This dataset has been referred from Kaggle. This is a latest dataset it's Uploaded 5 months ago.



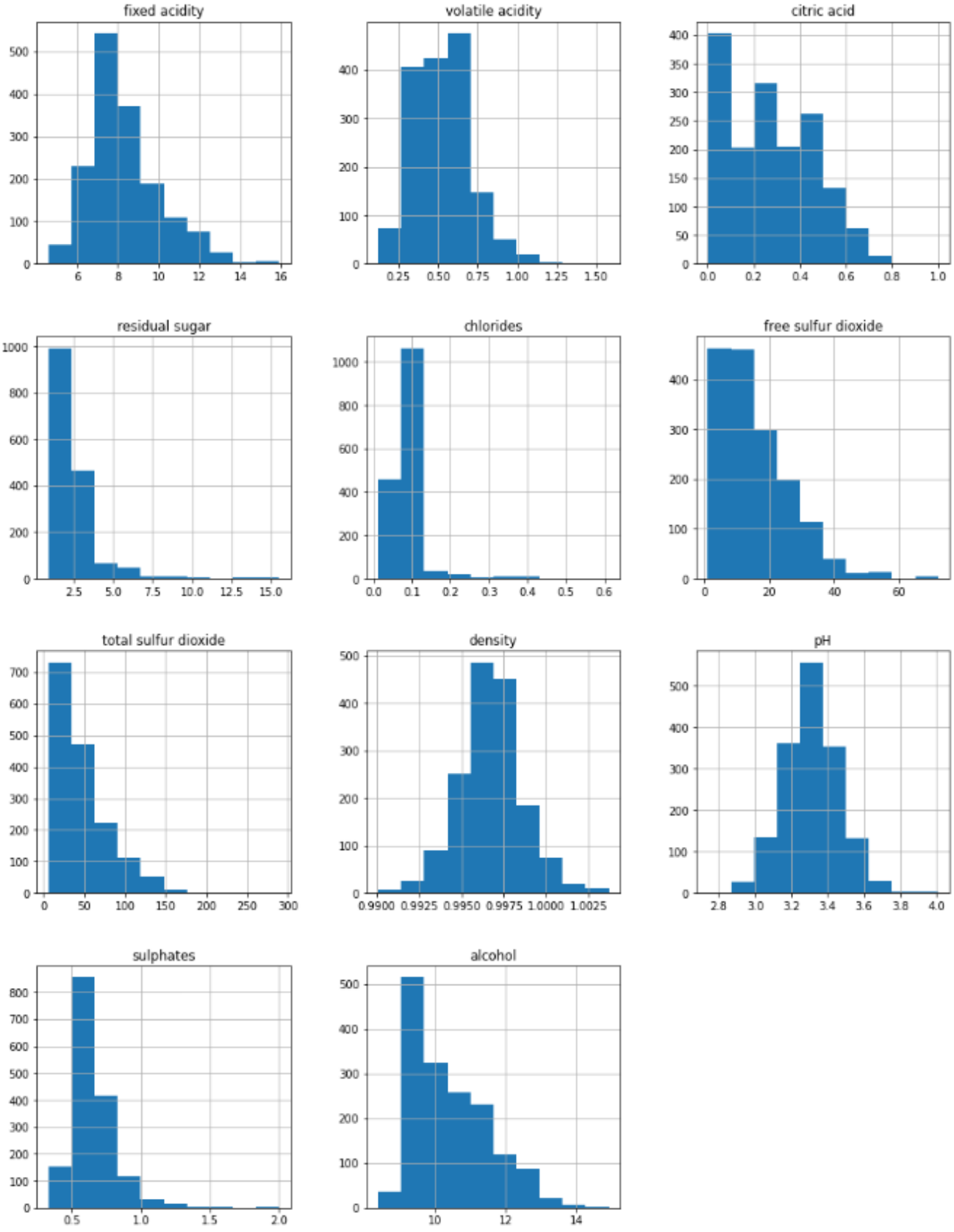
This dataset is also available from Kaggle & UCI machine learning repository

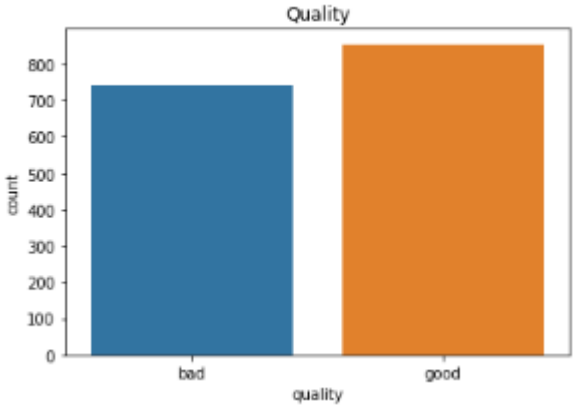
https://archive.ics.uci.edu/ml/datasets/wine+quality.

Methodology:

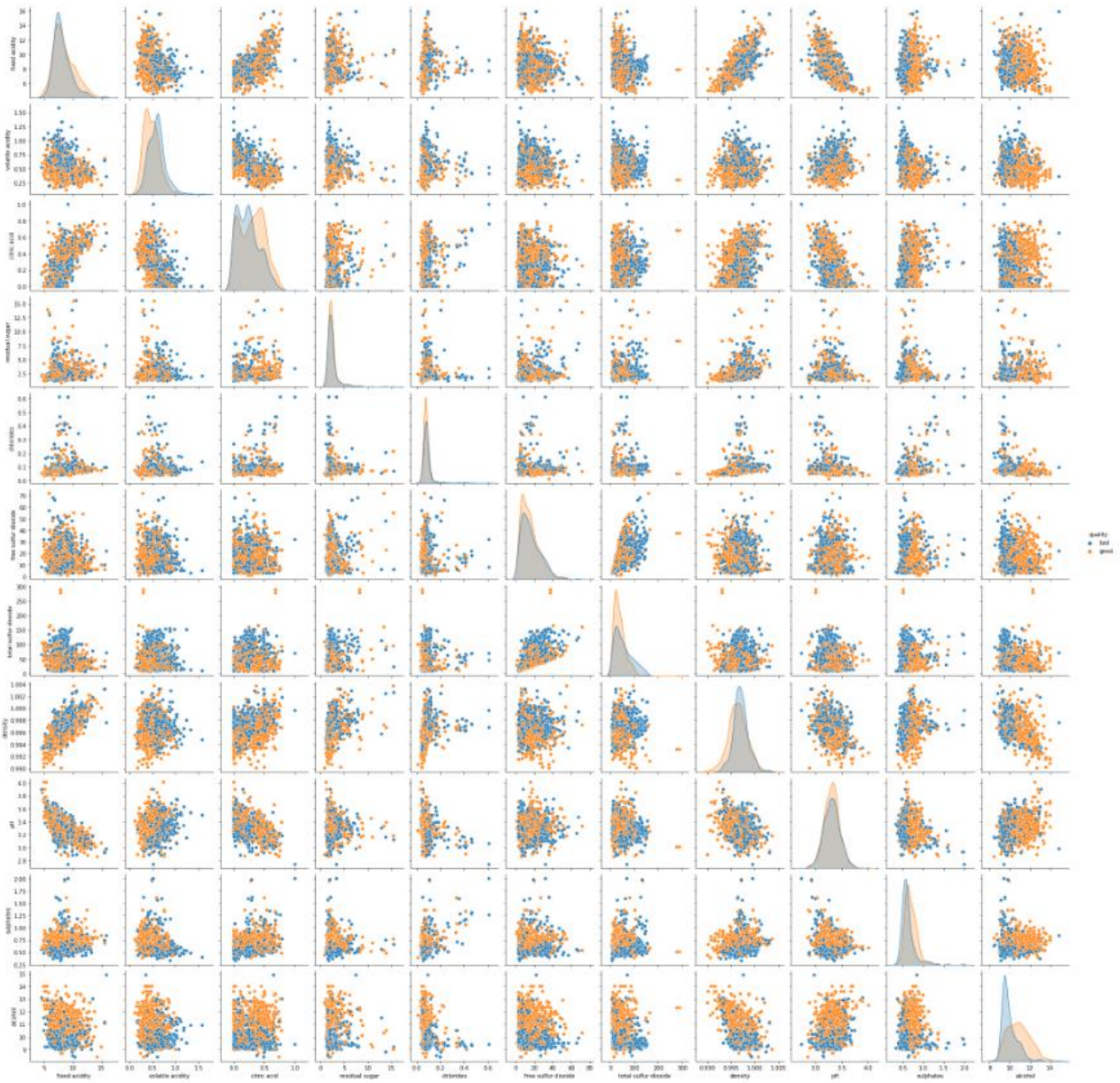
First, I mounted the dataset in Colab, then I checked for missing data, but there were no null values, therefore I removed the superfluous column. After dropping out, I created a new dataset, targeted the column, isolated the target column from the dataset, categorized the data set using four algorithms, and ultimately received result.

Block Diagram:

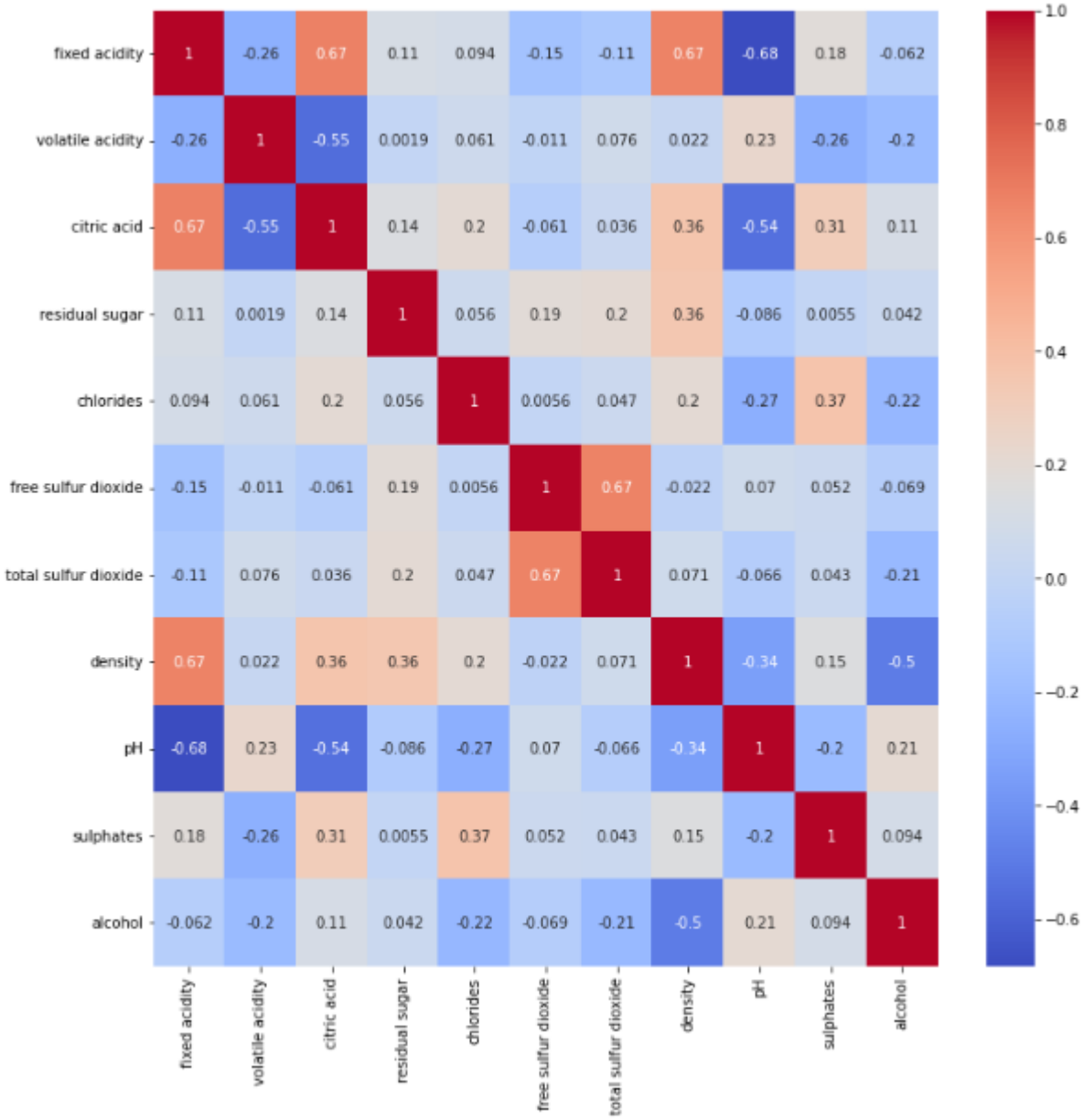




Data pairplot:



Heat map:



Algorithm:

Decision Tree, K-Nearest Neighbors, Random Forest, Logistic Regression

Result Analysis:

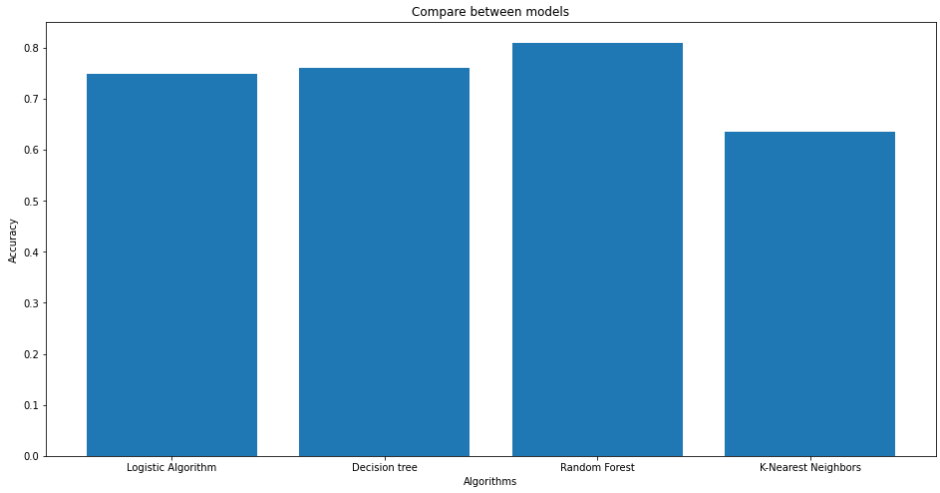
As a result, accuracy is frequently measured using the equation:

$$\text{Accuracy} = (TP+TN) / (TP+FP+FN+TN)$$
$$\text{Precision} = TP/TP+FP$$
$$\text{Recall} = TP/TP+FN$$
$$\text{F1 score} = TP/TP+1/2(FP+PN)$$

The accuracy values for all machine learning methods are shown in table:

No	Algorithm	Accuracy	Precision	Recall	F1 Score
1	Decision Tree	76%	76%	76%	76%
2	K-Nearest Neighbors	63%	64%	64%	64%
3	Random Forest	81%	81%	81%	81%
4	Logistic Regression	75%	75%	75%	75%

Compare Between Models:



Discussion:

We can see from the table above that Random Forest had the highest expected score of 81%, while KNN had the lowest predicted score of 63%. As a result, we can state categorically that Random Forest is superior since it has the highest projected score.

Conclusion:

This datasets is related to red variants of the Portuguese "Vinho Verde" wine.In this project in python, we learned to build a wine quality predictor on the csv dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. As a result, picking the relevant features and balancing the data in classification algorithms can increase the model's performance.