WINE QUALITY PREDICTION

CLASSIFICATION MACHINE LEARNING MODEL

PREPARED BY :- RAJ KUMAR

Table of content

ABSTRACT

Machine Learning is used across many ranges around the world. The food industry is no exception.Wines have been produced for thousands of years. But, it is a complex process to determine the relation between the subjective quality of a wine and its chemical composition.

Industries use Product Quality Certification to promote their products and become concerned for every individual who consumes any product. It is not possible to ensure quality with experts with such a huge demand for the product as it will increase the cost. Wine-makers need a permanent solution to optimize the quality of their wine. This paper explores the space to easy out and make the whole process cost-effective and more trustworthy using machine learning. aw

It allows firms to build a model with a user interface which predicts the wine quality by selecting the important parameters of wine which play a significant role in determining the wines quality. Our prediction model provides an ideal solution for the analysis of wine, which makes this whole process more efficient and cheaper with less human interaction.

INTRODUCTION

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increase their value in the market.

Previously, testing of product quality will be done at the end of the production, this is a time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion about the test, so identifying the quality of the wine based on human experts is a challenging task. There are several features to predict the wine quality but the entire features will not be relevant for better prediction. The research aims at what wine features are important to get the promising result by implementing the machine learning classification algorithms such as Logistic regression, Support Vector Machine (SVM), Naïve Bayes (NB), KNN-Classifier,Decision Tree, Random Forest classifier, using the wine quality dataset.

To evaluate our models, accuracy,classification report(precision, recall, and f1 score), Confusion matrix, are good indicators to evaluate the performance of the models.

METHODOLOGY

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts: training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

This system is implemented using the following modules.

1.)Collection of Dataset

2.)Data Preprocessing

3.)Model Building.

DATA COLLECTION

Initially, we collected a dataset for our Wine quality prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80% of training data is used and 20% of data is used for testing.

The dataset used for this project is WinequalityN https://www.kaggle.com/datasets/vishalkumbhar1997/wine-quality-prediction-with-logistic-regression is also available on Kaggle. The dataset consists of 13 attributes.

PRE- PROCESSING OF DATA

Data preprocessing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes.

In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

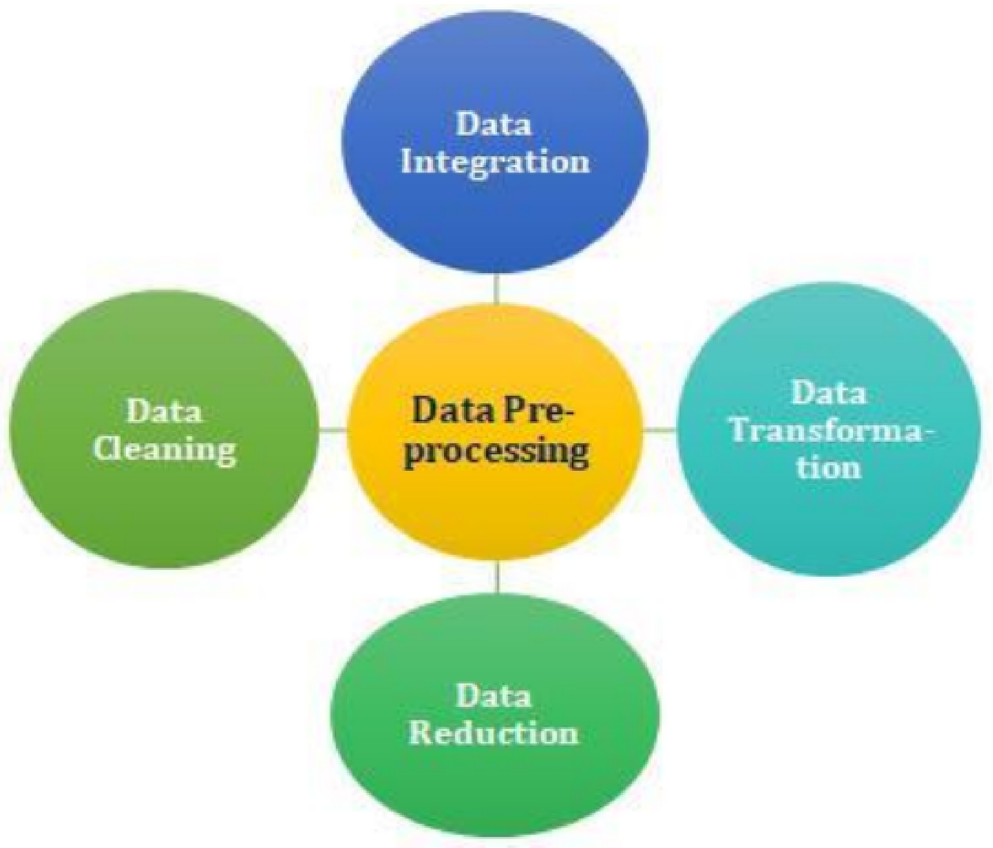
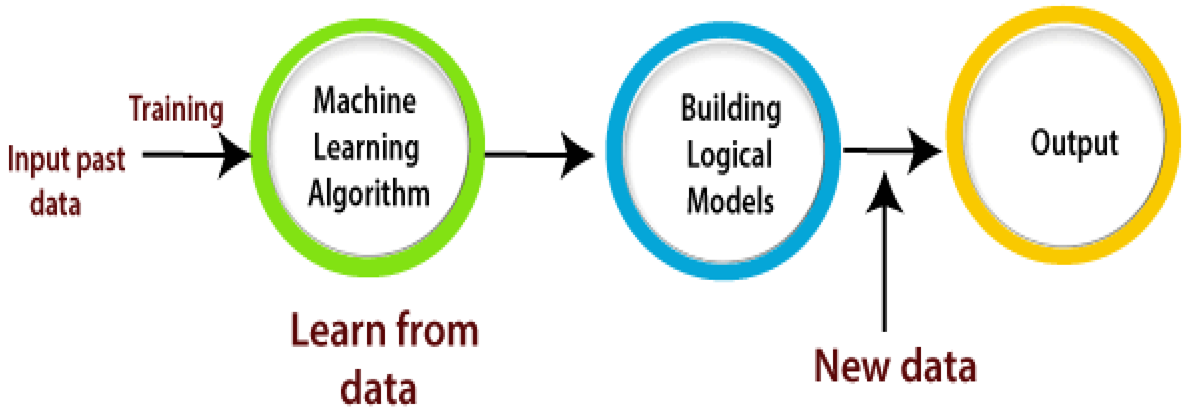


Figure: Data Pre-processing

MODEL BUILDING

Various machine learning algorithms like Logistic regression,SVM, Naive Bayes, Decision Tree, Random Tree are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the good outcome is used for wine quality prediction.

Performance metrics are accuracy score, classification report and confusion matrix to analyze models performance.

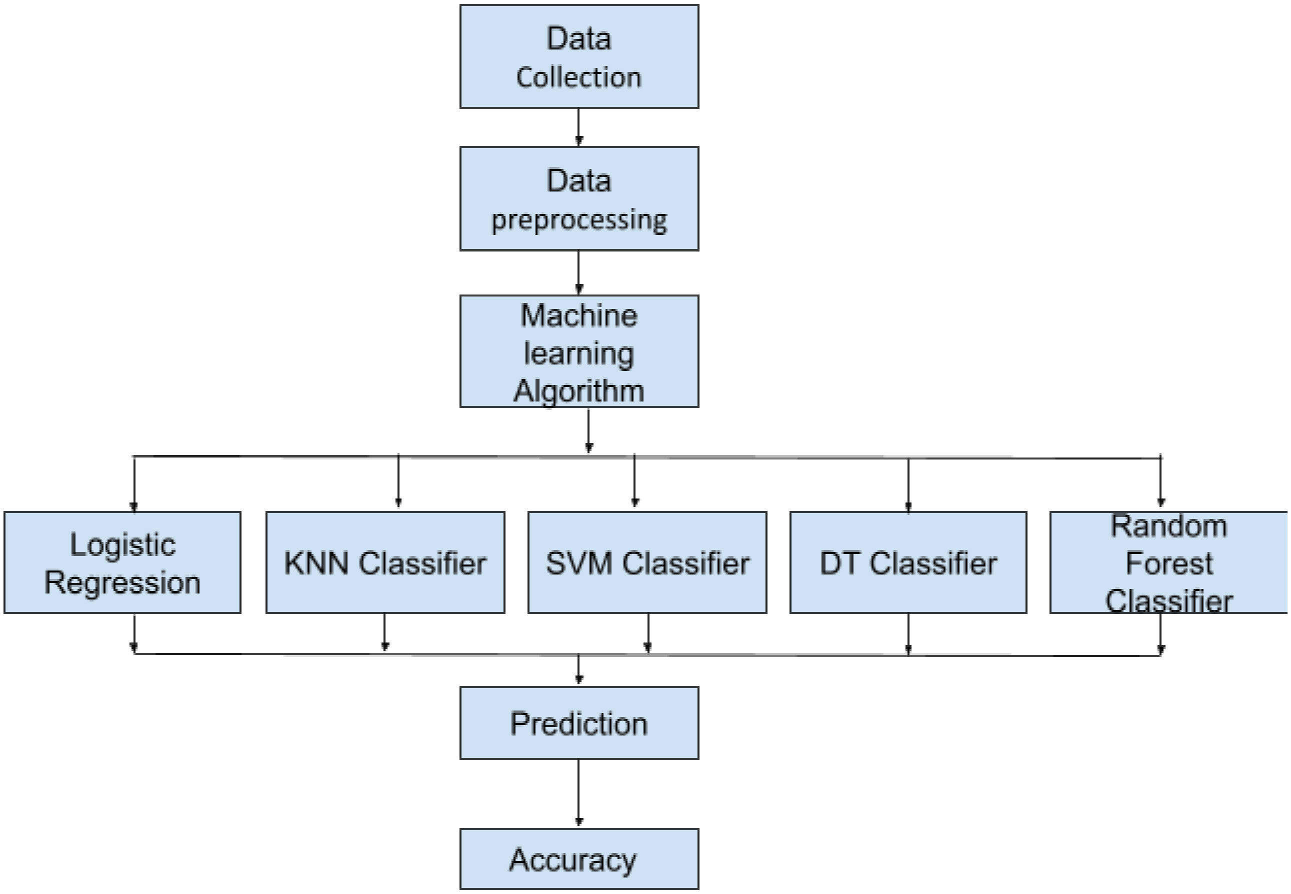


WORKING OF SYSTEM

SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system.

The working of this system is described as follows: -
Dataset collection is collecting data which contains wine details. Data is further cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of wine quality. Accuracy measure compares the accuracy of different classifiers. Other performance metrics are also considered for models.

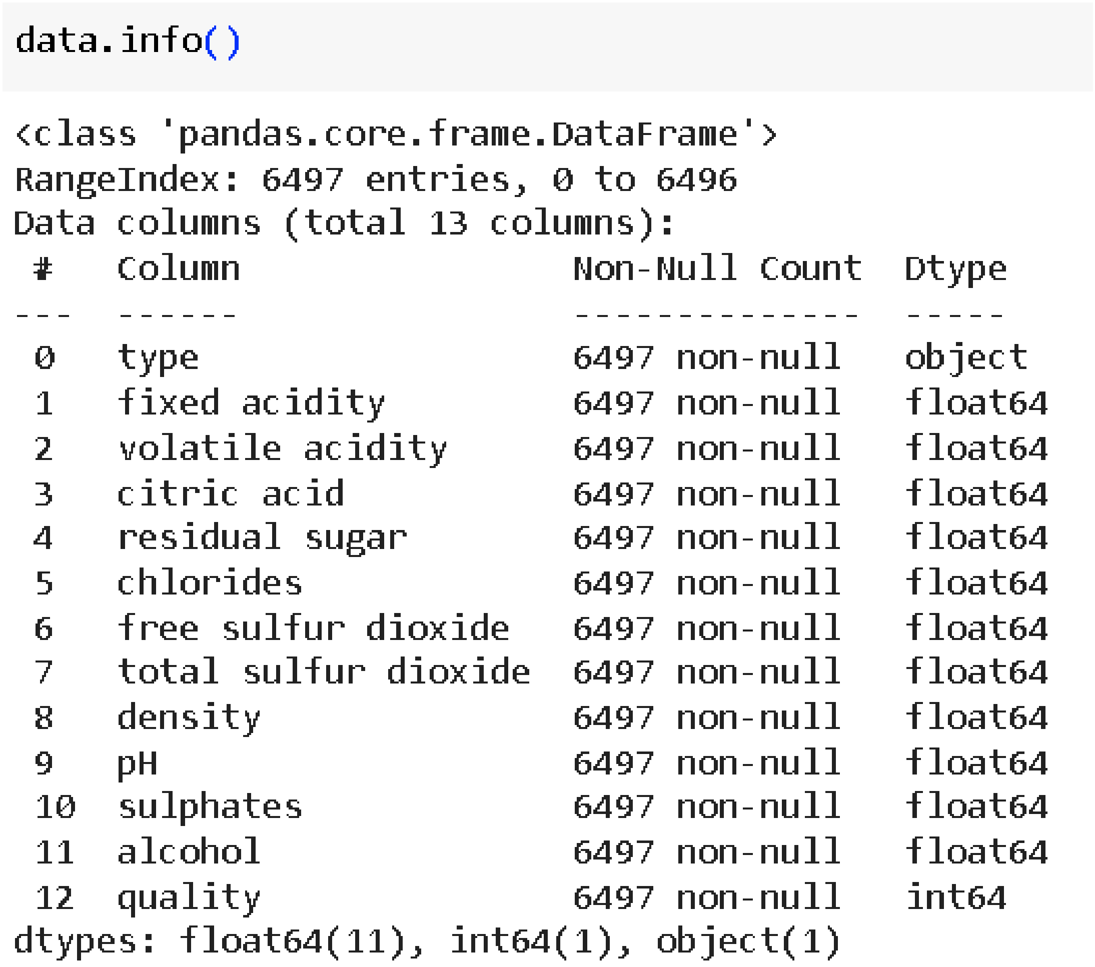


DATASET DESCRIPTION

Dataset contains following features:-

- Type:- Type of wine red/White
- 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 a 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:- Total sulfur dioxide
- pH:- In wine pH is used for checking acidity
- density
- sulfate:- Added sulfites preserve freshness and protect wine from oxidation, and bacteria.
- alcohol :- Percent of alcohol present in wine.
- Quality:- Quality of wine (rating given by users).



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   type                  6497 non-null   object 
 1   fixed acidity         6497 non-null   float64
 2   volatile acidity      6497 non-null   float64
 3   citric acid           6497 non-null   float64
 4   residual sugar        6497 non-null   float64
 5   chlorides             6497 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6497 non-null   float64
 10  sulphates             6497 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64  
dtypes: float64(11), int64(1), object(1)
```

From the dataset we can see we have 6497 rows and 13 features, in which 1 is object type, 1 is integer type and the rest 11 are float type data.

LIBRARIES USED

Python libraries make it easy for us to handle the data and perform typical and complex tasks with a single line of code.

Libraries used in this case study are as follow:-

- **Pandas**– Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

- **NumPy**–NumPy stands for Numerical Python. NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform, and matrices.

- **Matplotlib**– Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

- **Seaborn**– Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

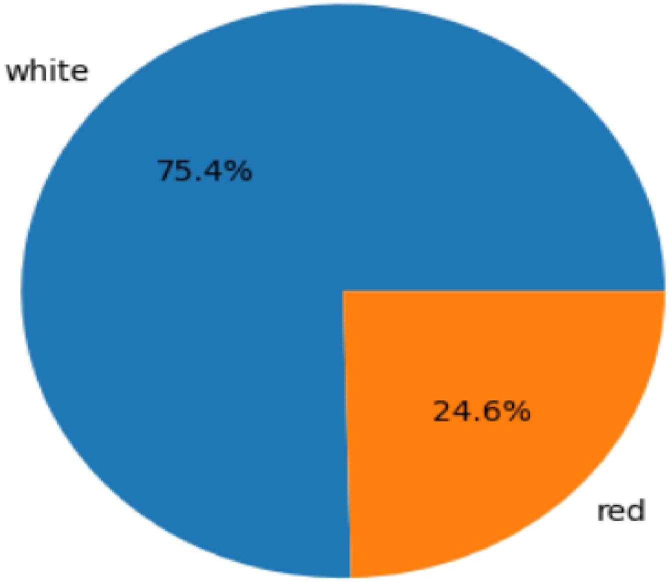
- **Sklearn**– This module contains multiple libraries having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.

EXPLORATORY DATA ANALYSIS

Univariate Analysis

TYPE

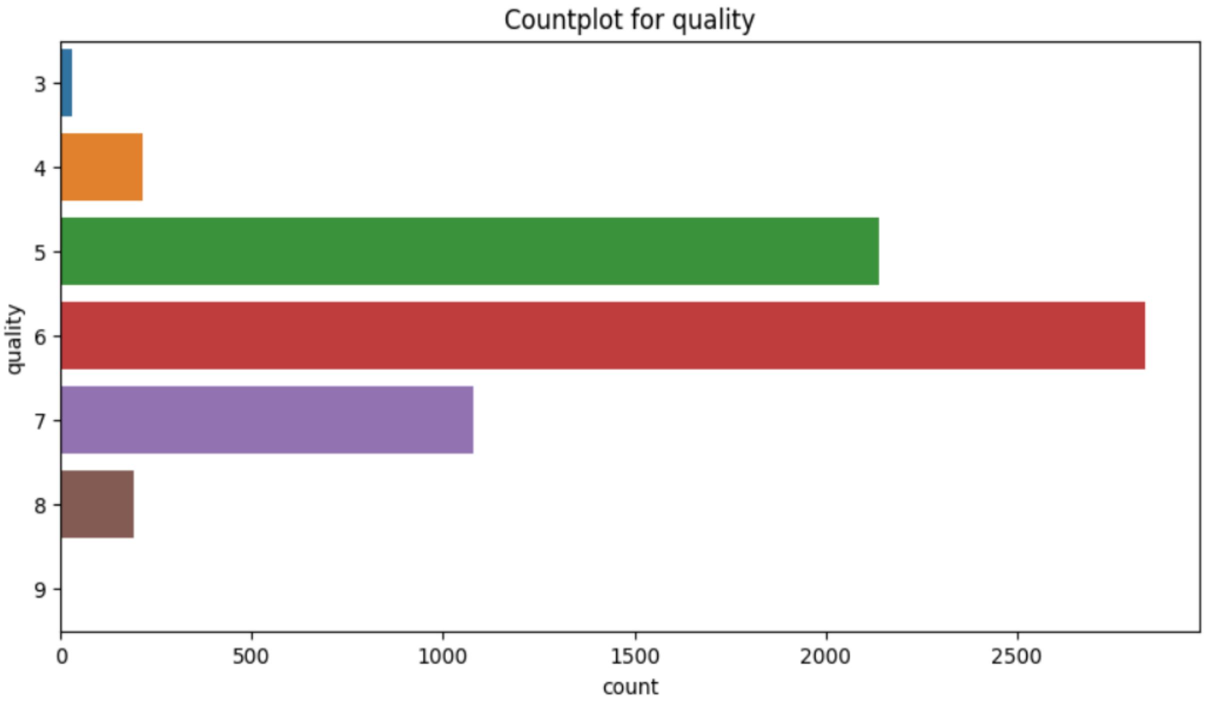
From the Pie-Chart distribution, we can see that the 75.4% samples are of White wines and 24.6% samples are for Red wines.



QUALITY

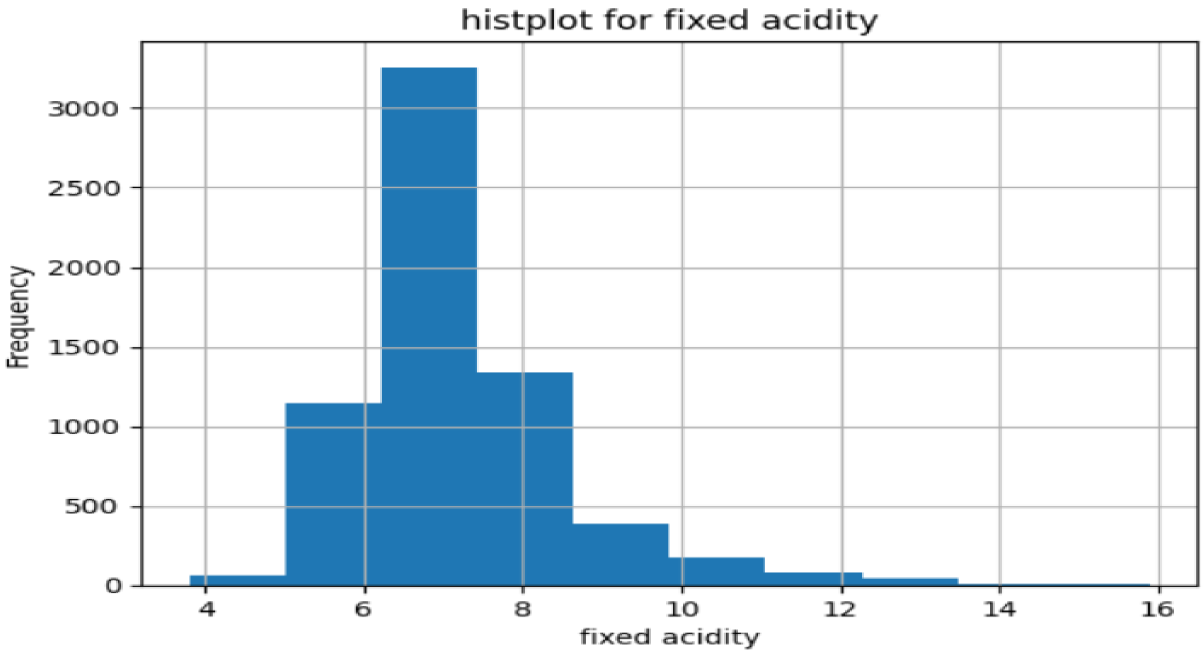
From the Countplot representation, we can see that the ratings given by users are ranging between 3 to 9. Here our aim is to predict the quality of wine.

We will create a new feature by using this quality feature. If the rating is more than 5, it will be treated as good, else it will be treated as bad quality.



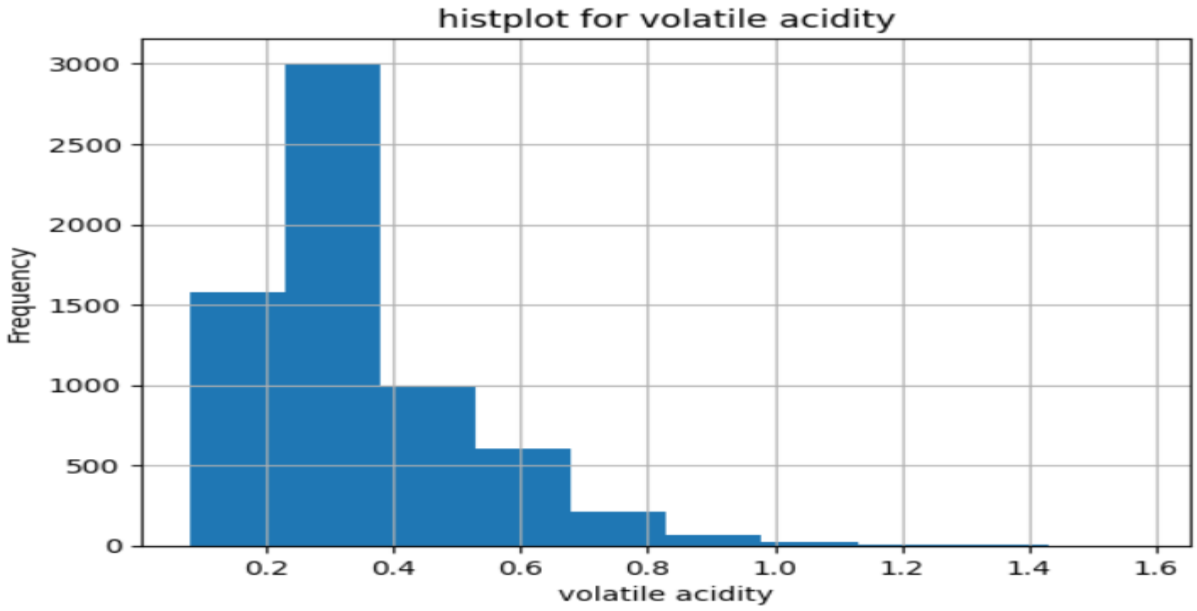
FIXED ACIDITY

Frequency distribution of Fixed acidity shows that most samples have acidity between 5 -9 units .



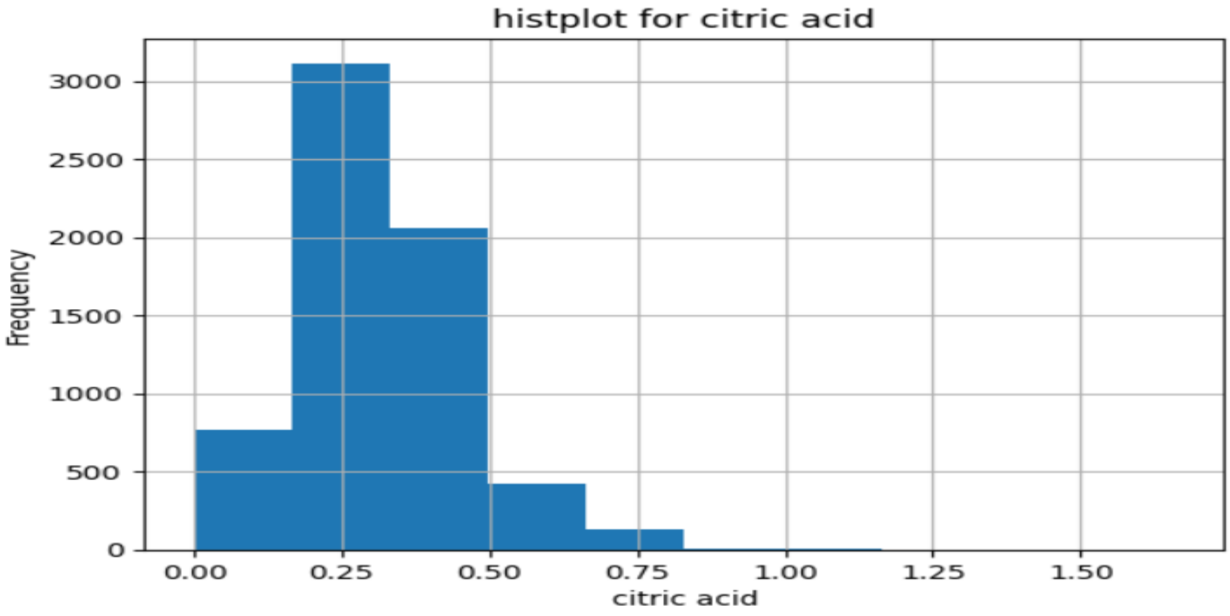
VOLATILE ACIDITY

From the distribution, we can see that there are many outliers in the dataset for Volatile acidity.



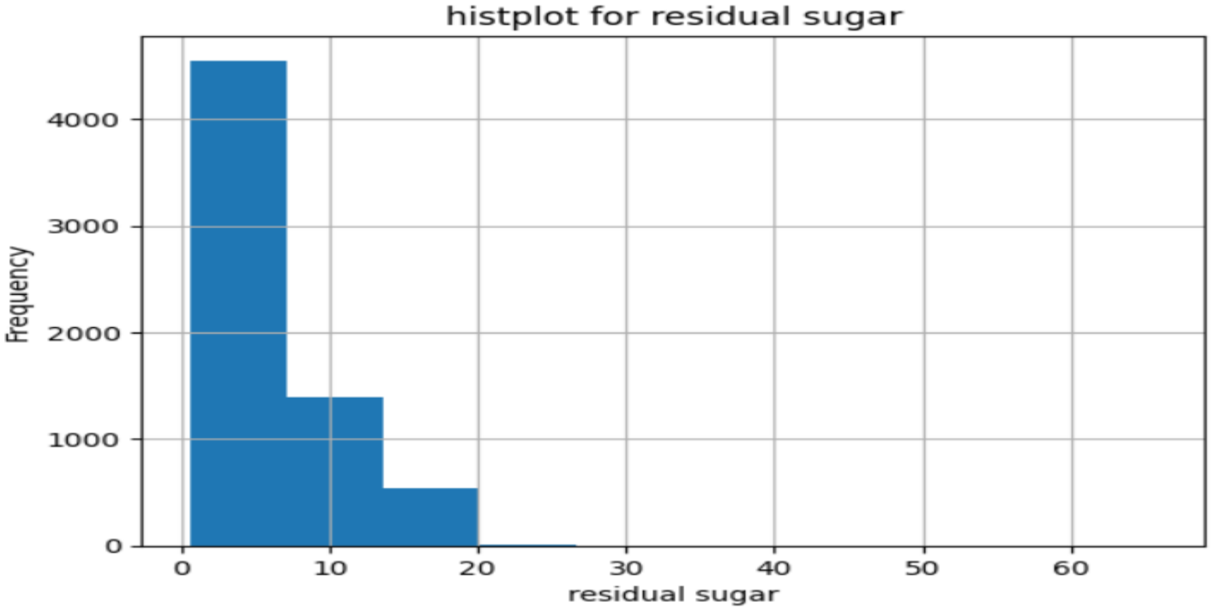
CITRIC ACID

Citric acid contains in range from 0 to 1.5 units, also there are some outliers in citric acid.



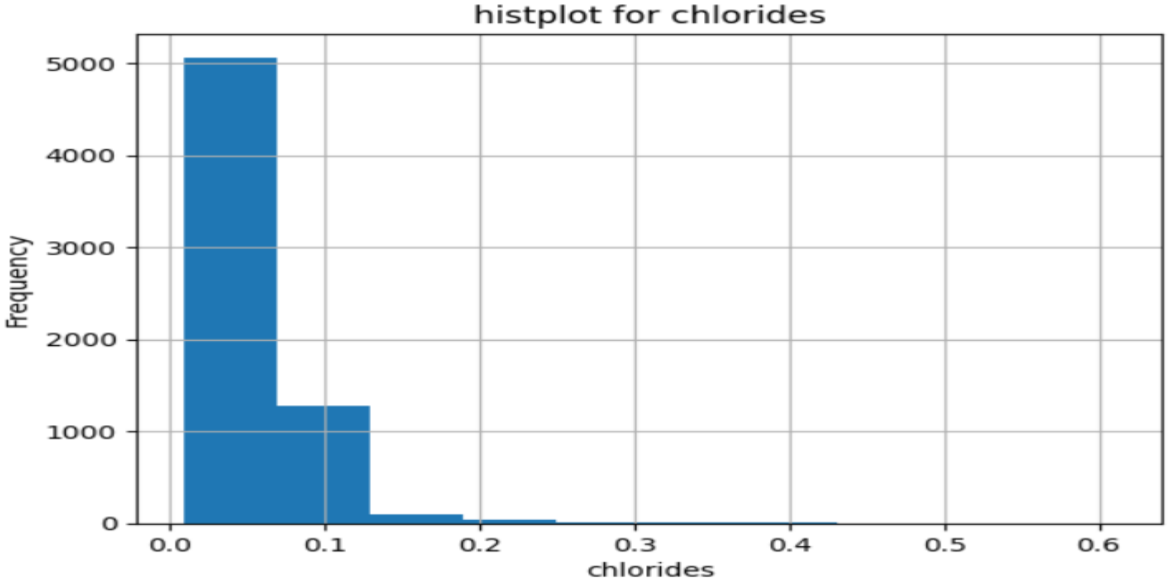
RESIDUAL SUGAR

From the frequency distribution, we can see that the residual sugar commonly persists up-to 20 units in wines .



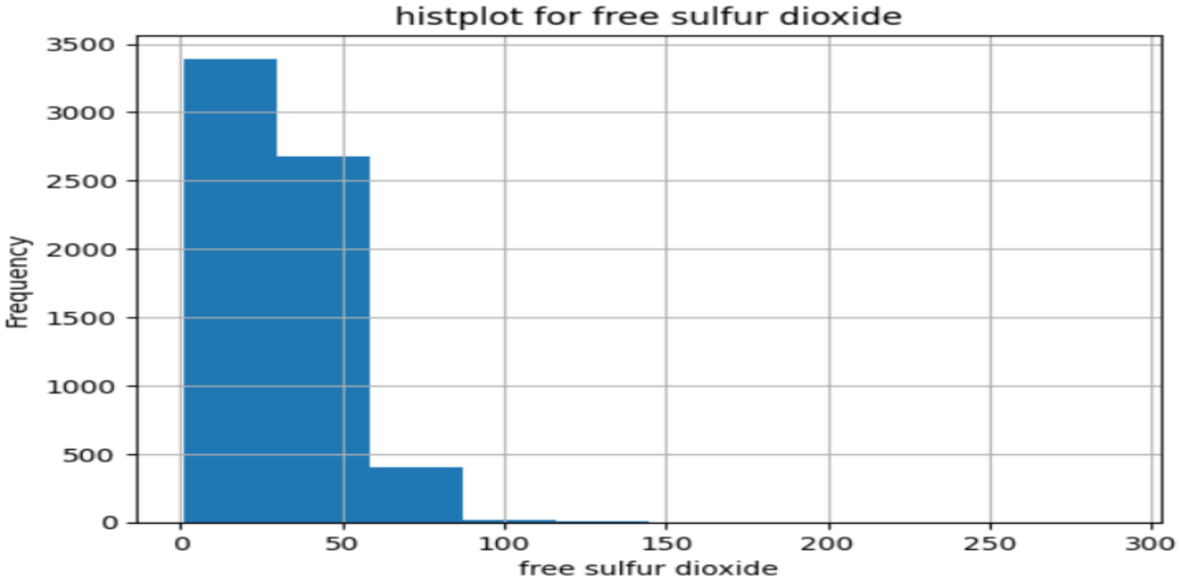
CHLORIDES

Most of the samples contain 0 to 0.15 unit Chlorides.



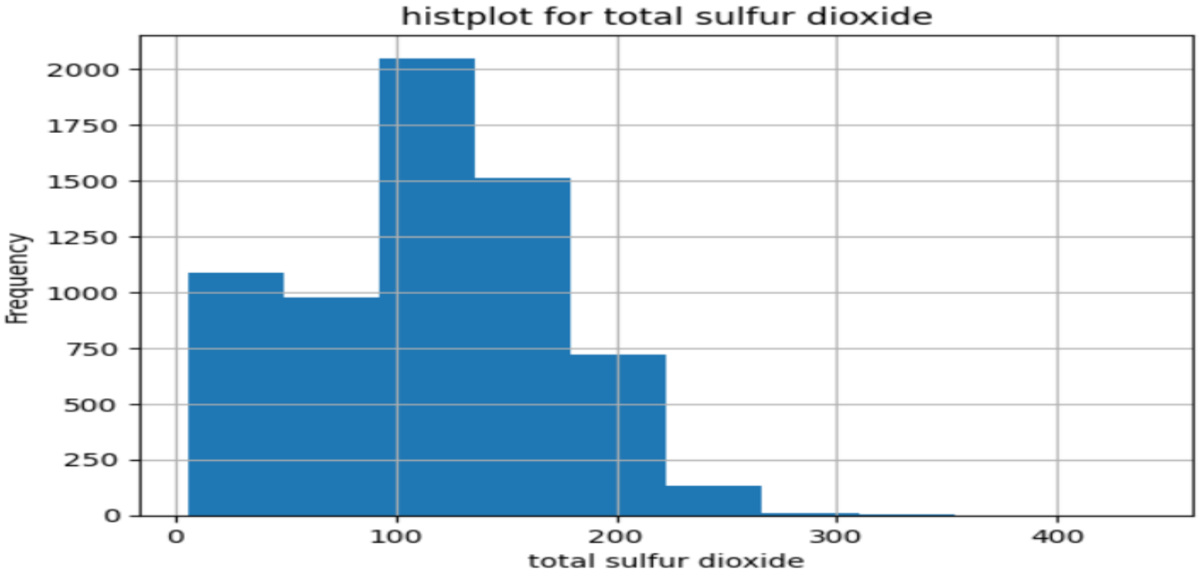
FREE SULFUR DIOXIDE

Most of the sample contains 0 to 50 unit Free sulfur dioxide.



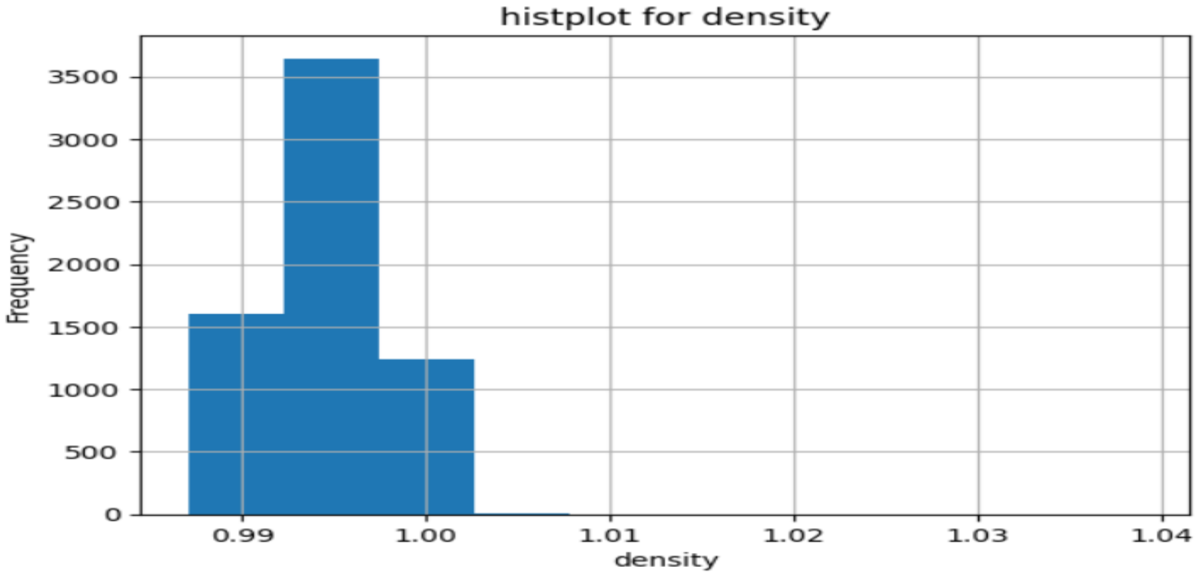
TOTAL SULFUR DIOXIDE

Most of the sample contains 0 to 200 unit Free sulfur dioxide.



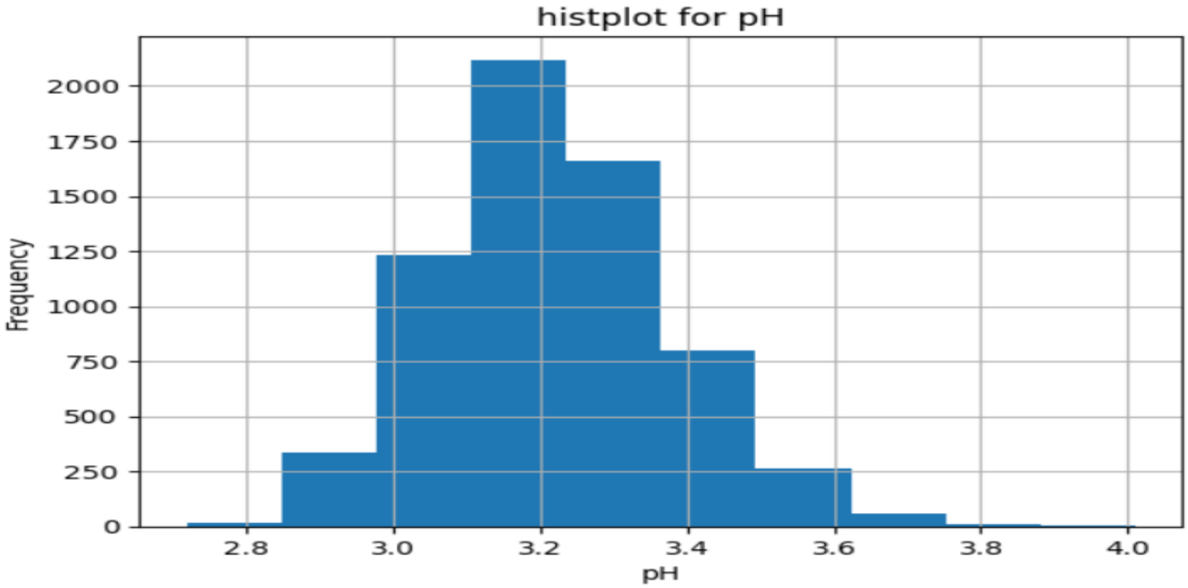
DENSITY

Density of liquid/ wine is almost 1%.



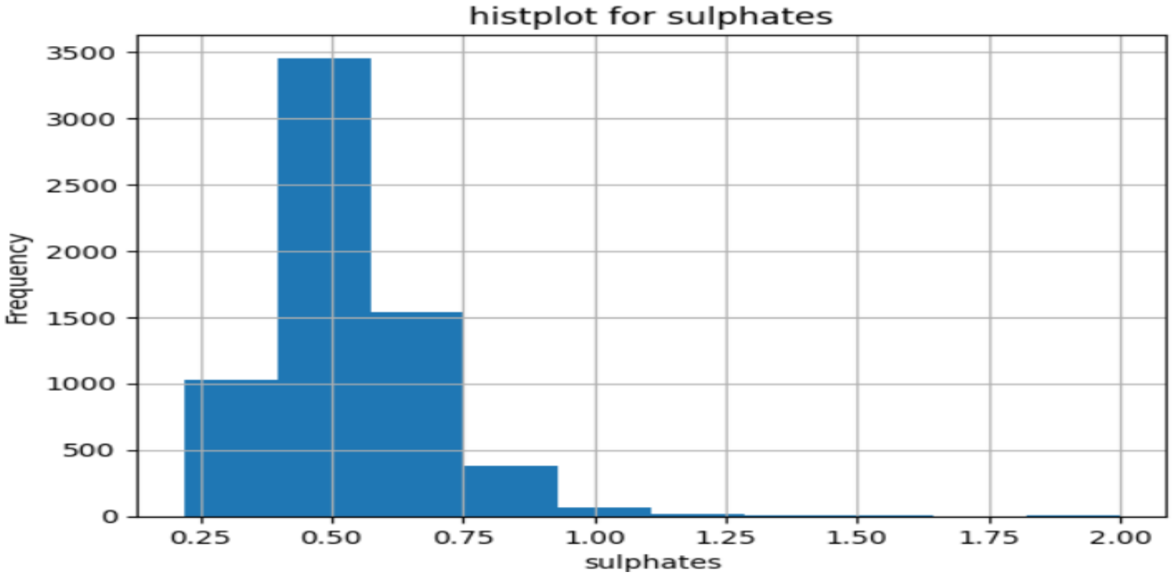
PH

The PH value of most of the wines are between 2.8 and 3.6.



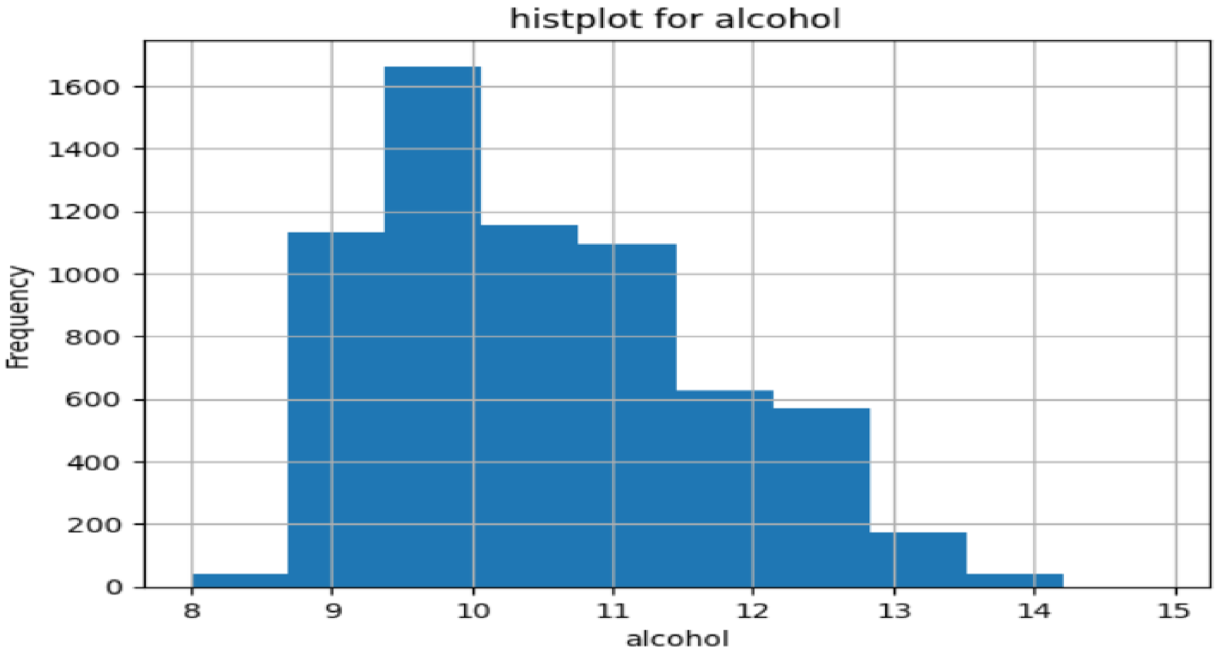
SULPHATES

Most of the sample lies in the range of 0.25 to 0.75 for sulphates..



ALCOHOL

From the frequency distribution, we can see that 9 to 13 % alcohol is contained in maximum wines.



MODEL BUILDING AND EVALUATION

We will use this wine quality dataset for training and testing in the ratio of 80-20. Model will train using 80% of data and evaluate on the rest 20% dataset.
We will evaluate model performance with different algorithms and also check performance after fine tuning of all models.

EVALUATION

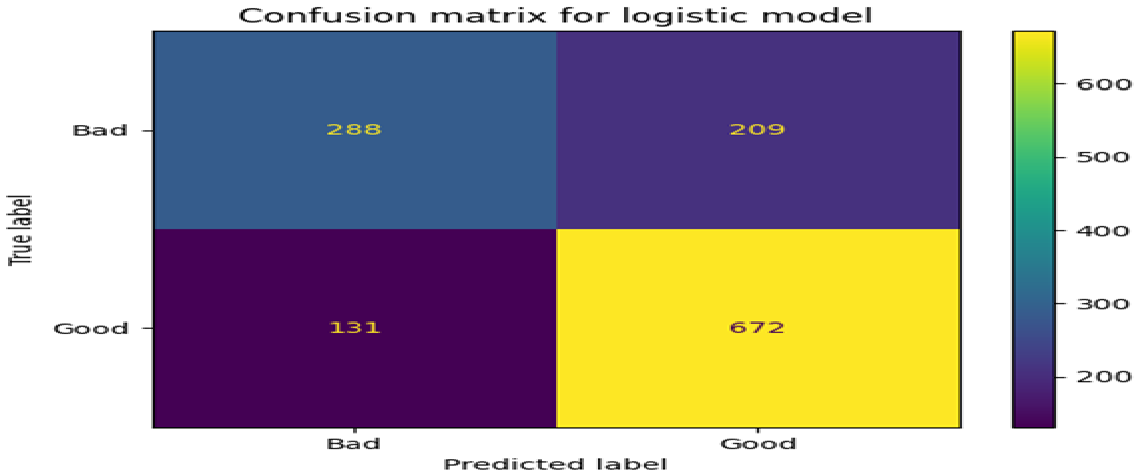
To evaluate our model we will use accuracy score and confusion matrix as performance metrics. We can't evaluate only based on accuracy score as our dataset is imbalanced.

LOGISTIC REGRESSION MODEL

ACCURACY SCORE

- Accuracy score for training data is : 0.7465845680200115
- Accuracy score for test data is : 0.7384615384615385

CONFUSION MATRIX

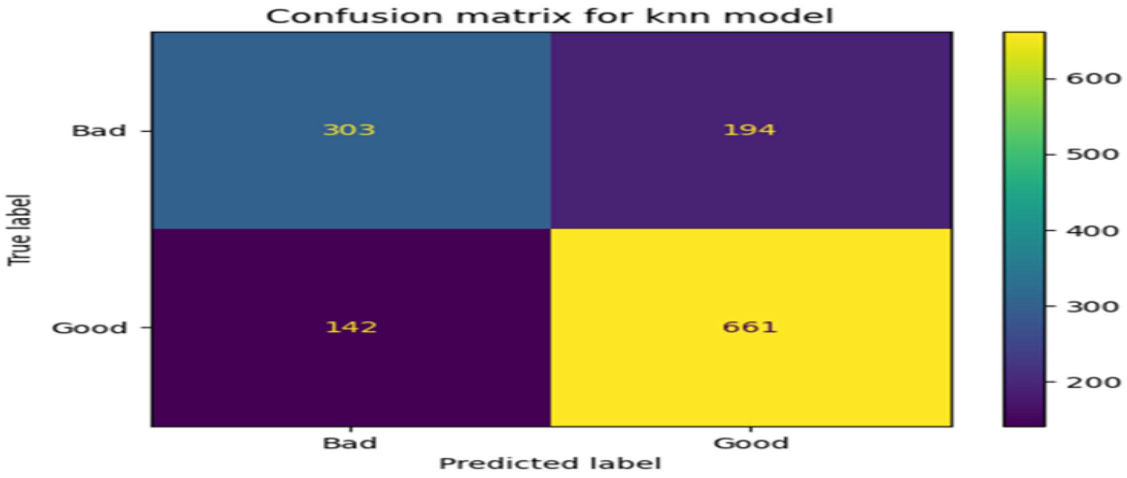


KNN MODEL

ACCURACY SCORE

- Accuracy score for train data is : 0.8379834519915336
- Accuracy score for test data is : 0.7415384615384616
- Train accuracy is much more than Test Accuracy i.e. slight overfitting is there.

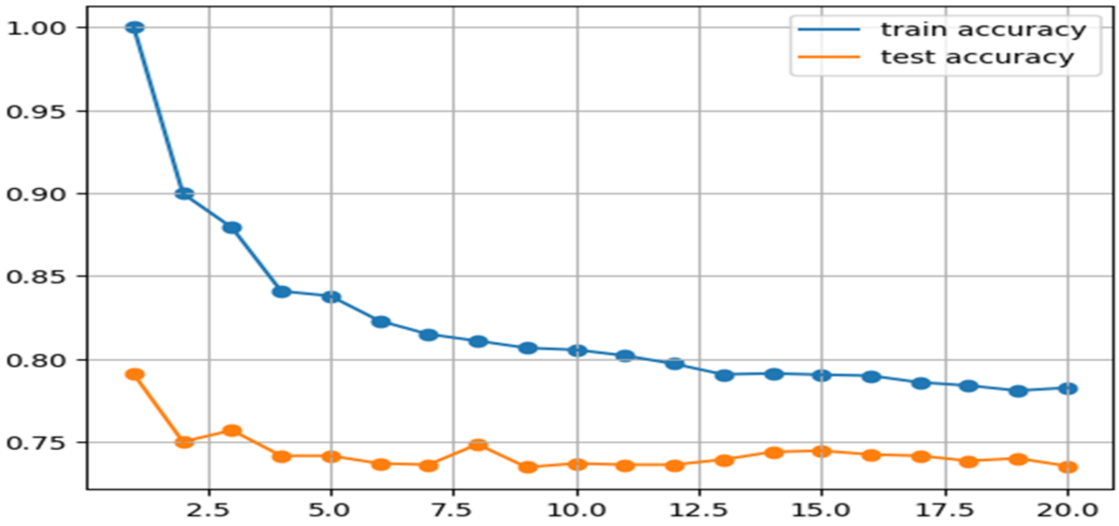
CONFUSION MATRIX



As KNN models predict output based on the value of K selected for training. At a point model try to fit train and test accuracy at a point.

We can tune the model for different no. of K to achieve high accuracy and overcome underfitting and overfitting.

In this model we will check for accuracy on different values of k from 1 to 2



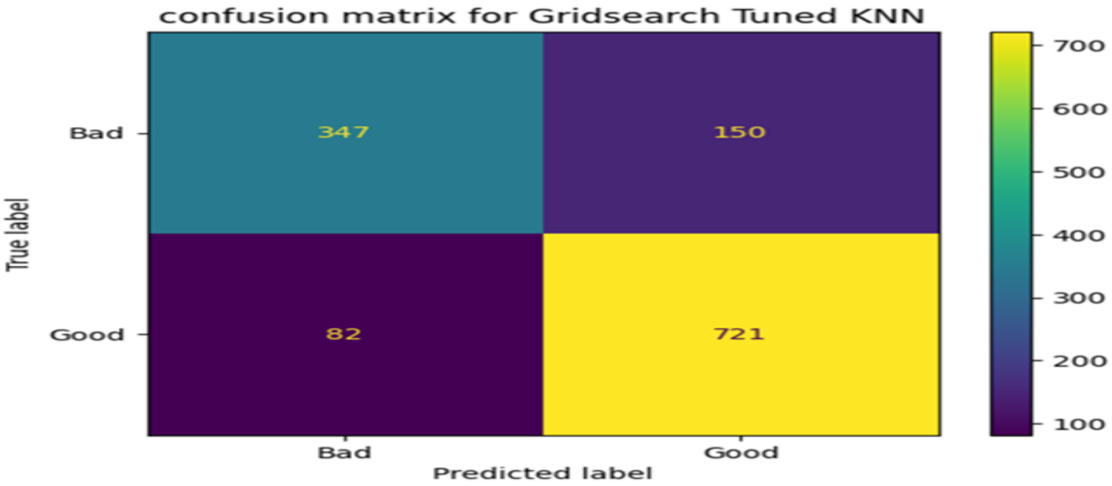
As we can see train and test accuracy trying to match at a point, on the basis of the above graph we can select k=8 for best performance.

TUNED KNN

ACCURACY SCORE

- Accuracy score for tuned KNN is : 0.8215384615384616
- Accuracy score improved from 74% to 82% By tuning.

CONFUSION MATRIX



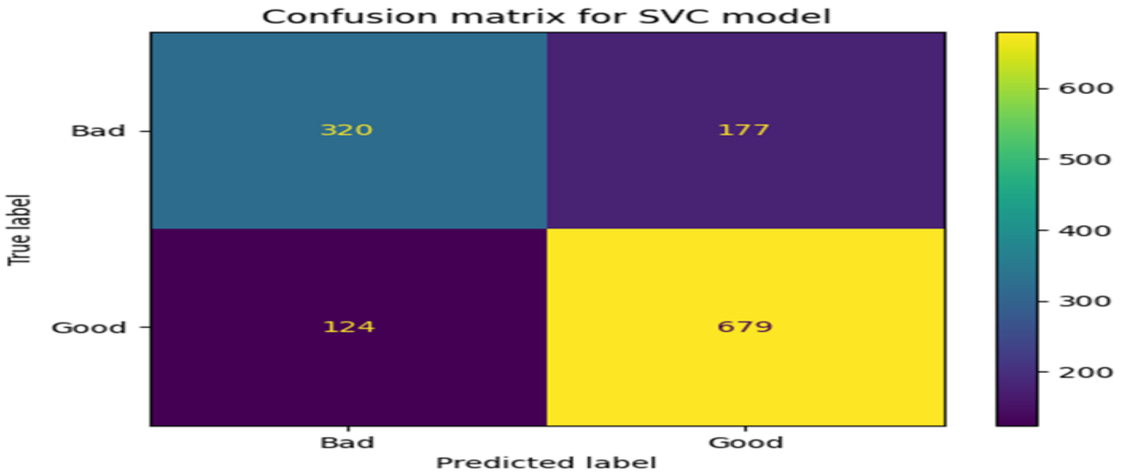
SUPPORT VECTOR CLASSIFIER MODEL

ACCURACY SCORE

Accuracy score for train data is : 0.8035405041370021

Accuracy score for test data is : 0.7684615384615384

CONFUSION MATRIX

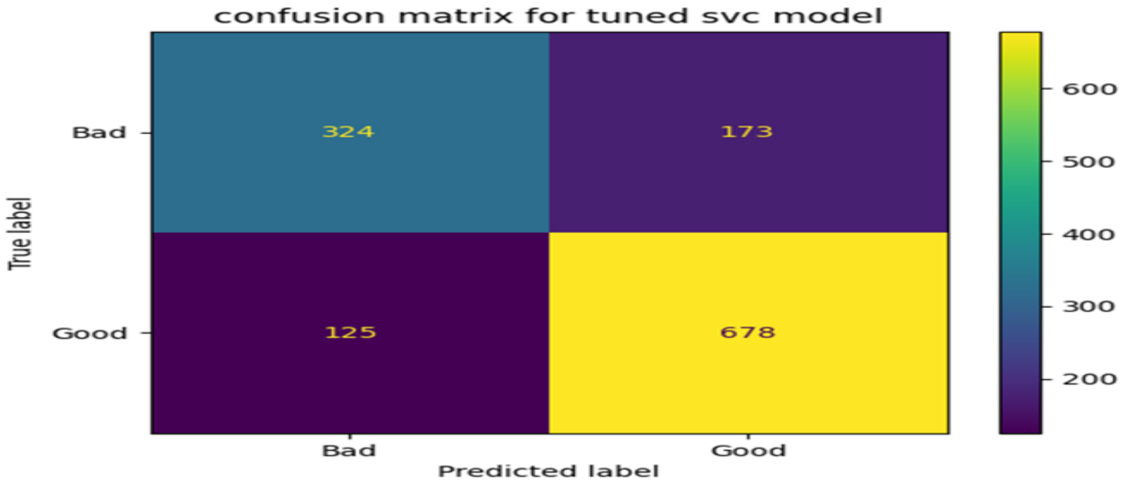


TUNED SUPPORT VECTOR CLASSIFIER MODEL

ACCURACY SCORE

- Accuracy score for Tuned SVC model is : 0.7707692307692308

CONFUSION MATRIX



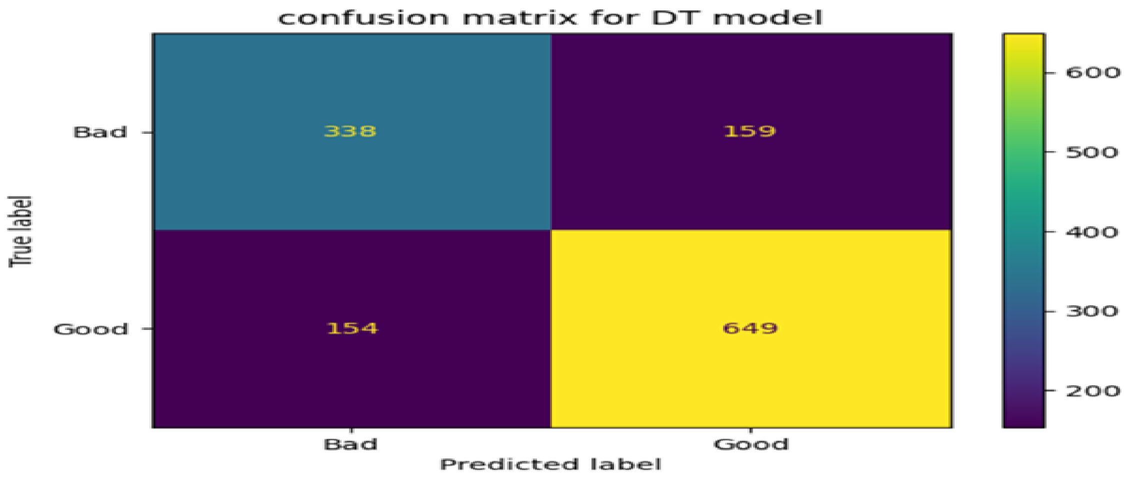
DECISION TREE CLASSIFIER MODEL

ACCURACY SCORE

Accuracy score for train data is : 1.0

Accuracy score for test data is : 0.75923076923076

CONFUSION MATRIX

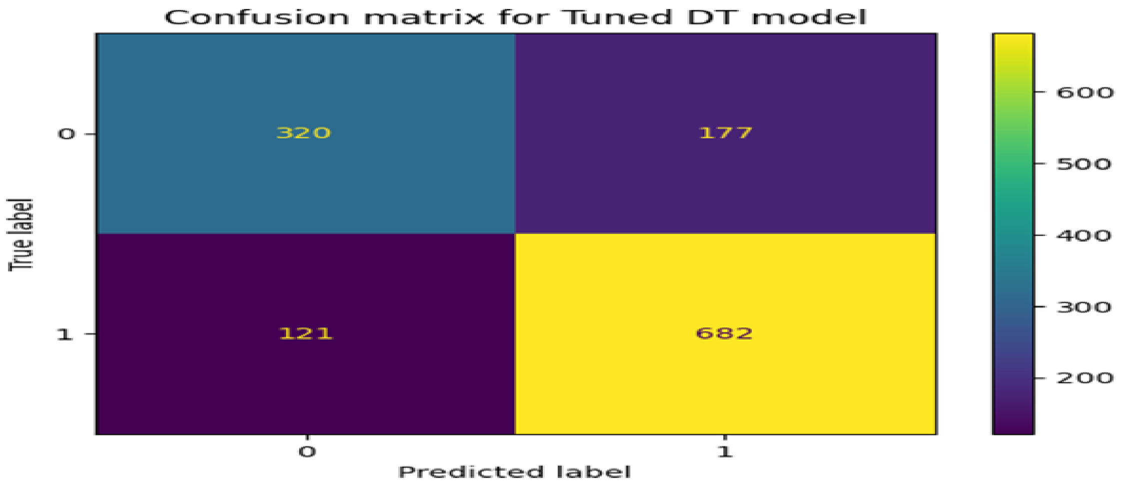


TUNED DECISION TREE CLASSIFIER MODEL

ACCURACY SCORE

- Accuracy score for Tuned Decision tree classifier model is :0.7707692307692

CONFUSION MATRIX



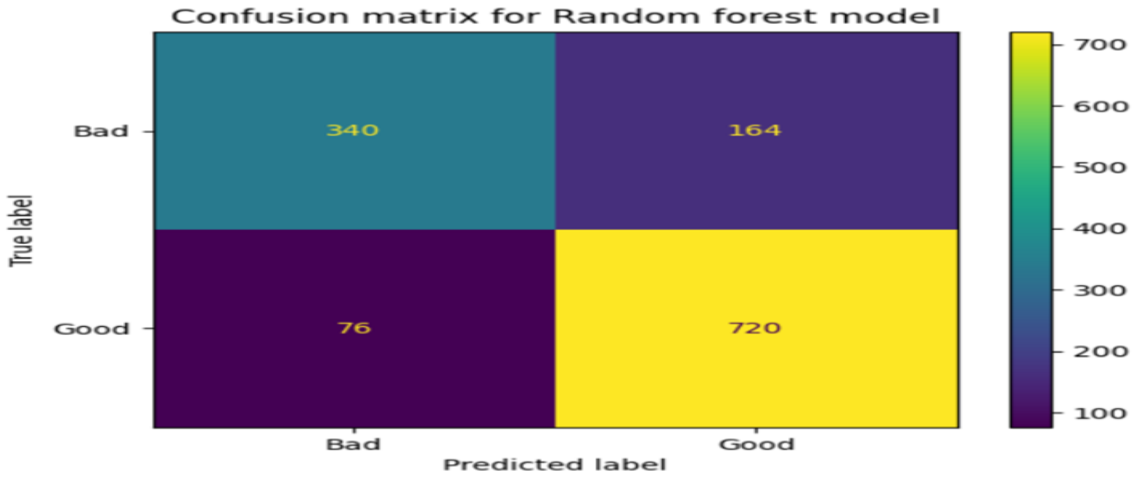
RANDOM FOREST CLASSIFIER MODEL

ACCURACY SCORE

Accuracy score for train data is : 0.999615

Accuracy score for test data is : 0.815323

CONFUSION MATRIX

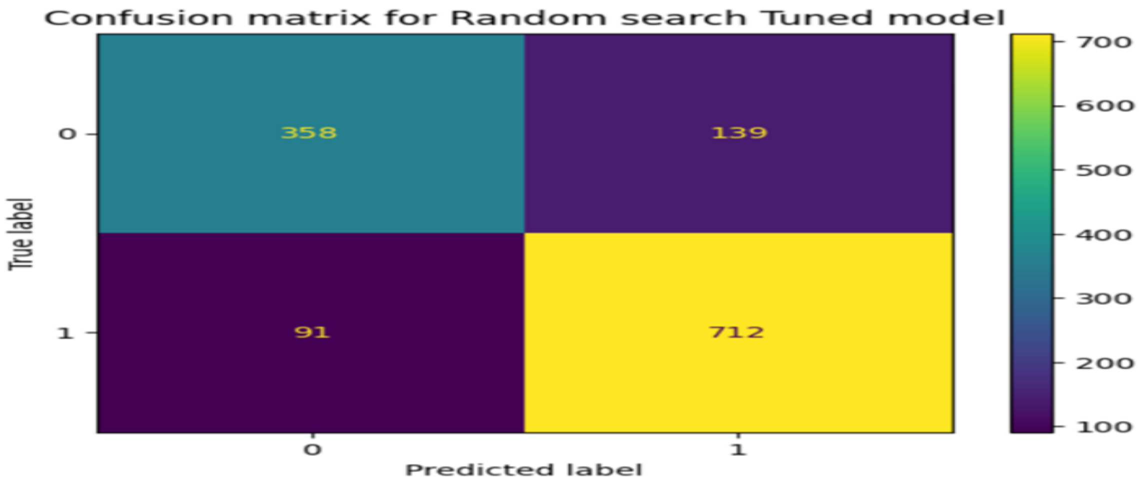


TUNED RANDOM FOREST CLASSIFIER MODEL

ACCURACY SCORE

- Accuracy score for Tuned Decision tree classifier model is :0.82307692

CONFUSION MATRIX



CONCLUSION

We started with the data preprocessing where we checked for missing data,filled data for missing values, and created new features. Afterward, we perform EDA on the dataset to explore all the features . Then after EDA started training on different machine learning models, and for each model, we did hyperparameter tuning and checked with the Accuracy score and also with a confusion matrix.

Based on prediction we conclude -

- □ Based on accuracy score we can see that KNN model is more accurate to predictions after tuning.
- □ Random forest model is also good with both categories But after tuning it is performing good for only one category.
- □ Slight improvement in Confusion matrix and Accuracy score is observed in models.

FUTURE SCOPE OF IMPROVEMENTS

As we can see the prediction on models , Some models are performing very well with training data but not able to retain the same performance with test/unseen data.While training, loss also occurs in some models. Hence, The model is overfitted.

To improve the performance in future we can do some more preprocessing and select only relevant features for model training, we can add more models as well to get better results.

REFERENCES

https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html