**WINE QUALITY PREDICTION**

# TERM PROJECT

***Submitted by***

# CB.EN.U4CCE21009 ASHWANTHRAM AC

**CB.EN.U4CCE21015 DHANUSHRAJU S**

# CB.EN.U4CCE21020 GAUTAM RS

# CB.EN.U4CCE21026 HARINI RAJ AKHTSHYA RG



**B.TECH COMPUTER AND COMMUNICATION ENGINEERING**

# 2021 BATCH

**PAPER SUMMARY**

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 increases their value in the market. Previously, testing of product quality will be done at the end of the production, this is 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 humans experts it 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 to what wine features are important to get the promising result by implementing the machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), using the wine quality dataset.The algorithms, LR, RF, NB, DT ,SVM and NB give us the cross validation accuracy of approximated to 73.1%, 72.1%, 71.7%, 63.2%, 63% and 59.2%. Amongst all Random Forest Algorithm (RF) gives us the best accuracy of 74% and with this algorithm we can predicted whether the quality of the wine is good or bad

**METHODOLOGY**

In this method, we begin by acquiring a dataset and subsequently utilize machine learning methods to preprocess the data, replace the missing values by mean and also finding out the outliers and replacing it by its mean. This helps in simplifying the data by reducing its complexity. Next, we apply Logistic Regression Algorithm, Decision Tree, Random Forest Algorithm, KNN Classifier, Support Vector Machine and Naive Bayes Classification algorithms to the processed dataset. By considering the characteristics of the dataset, we make predictions. And we compare all the algorithms and get their accuracy level. With this we come to a conclusion that,which algorithm suits the best for this model. And the predictions inform us whether the quality of thw wine is good or bad.

**1) DATASET DESCRIPTION**

The wine dataset contains 1599 instances and. The dataset have 11 input variables (based on physicochemical tests): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and 1 output variable (based on sensory data): quality. Sensory data is scaled in 11 quality classes from 0 to 10 (0-6.5 says the quality of wine is bad and the value with above 6.5 says the quality of the wine is good).

|  |  |
| --- | --- |
| **ATTRIBUTE** | **DESCRIPTION** |
| **Fixed acidity** | **Fixed acids, numeric from 3.8 to 15.9** |
| **Volatile acidity** | **Volatile acids, numeric from 0.1 to 1.6** |
| **Citric acid** | **Citric acids, numeric from 0.0 to 1.7** |
| **Residual sugar** | **residual sugar, numeric from 0.6 to 65.8** |
| **Chlorides** | **Chloride, numeric from 0.01 to 0.61** |
| **Free sulfur dioxide** | **Free sulfur dioxide, numeric: from 1 to 289** |
| **Total sulfur dioxide** | **Total sulfur dioxide, numeric: from 6 to 440** |
| **Density** | **Density, numeric: from 0.987 to 1.039** |
| **pH** | **pH, numeric: from 2.7 to 4.0** |
| **Sulfates** | **Sulfates, numeric: from 0.2 to 2.0** |
| **Alcohol** | **Alcohol, numeric: from 8.0 to 14.9** |
| **Quality** | **Quality, numeric: from 0 to 10, the output target** |

**2) DATASET PREPROCESSING**

Data pre-processing includes defining independent variables and response variable,

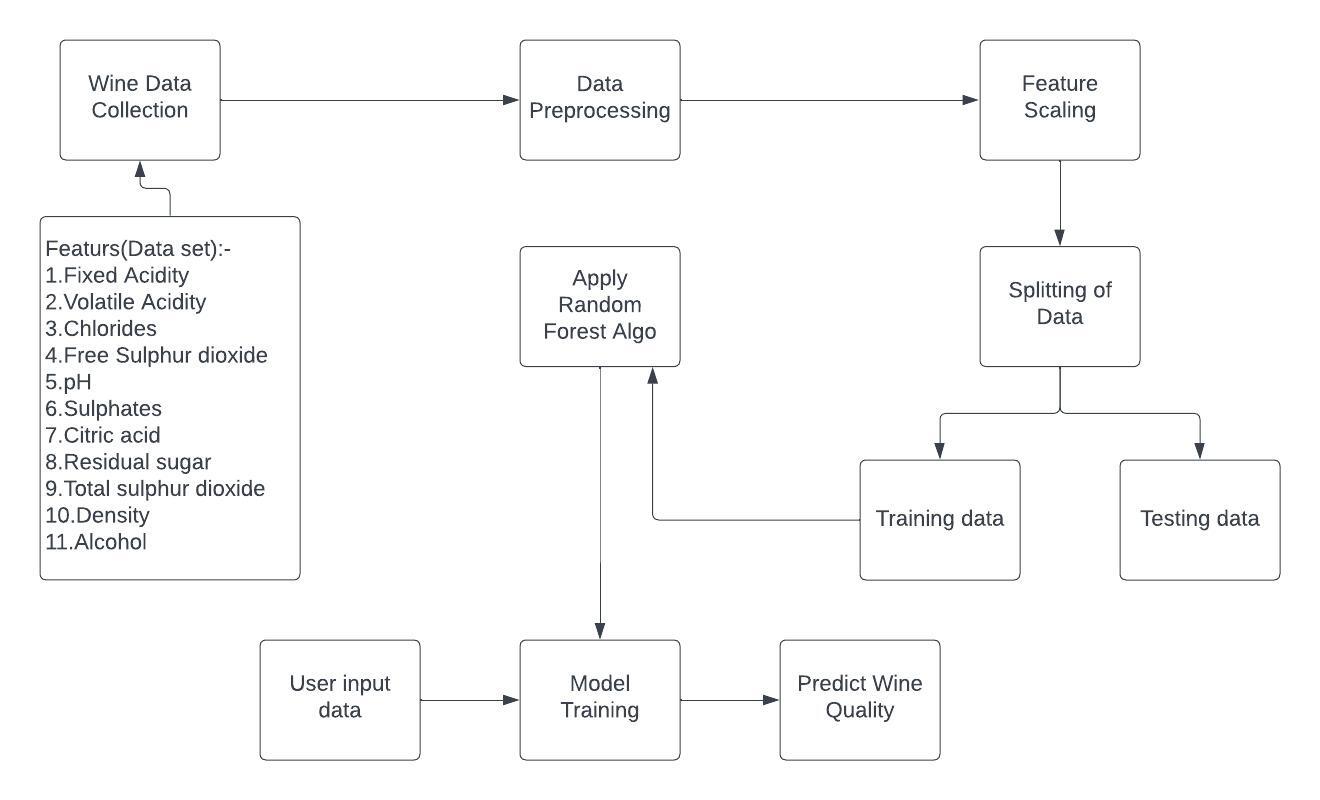
missing data handling, feature scaling and data splitting into train data and test

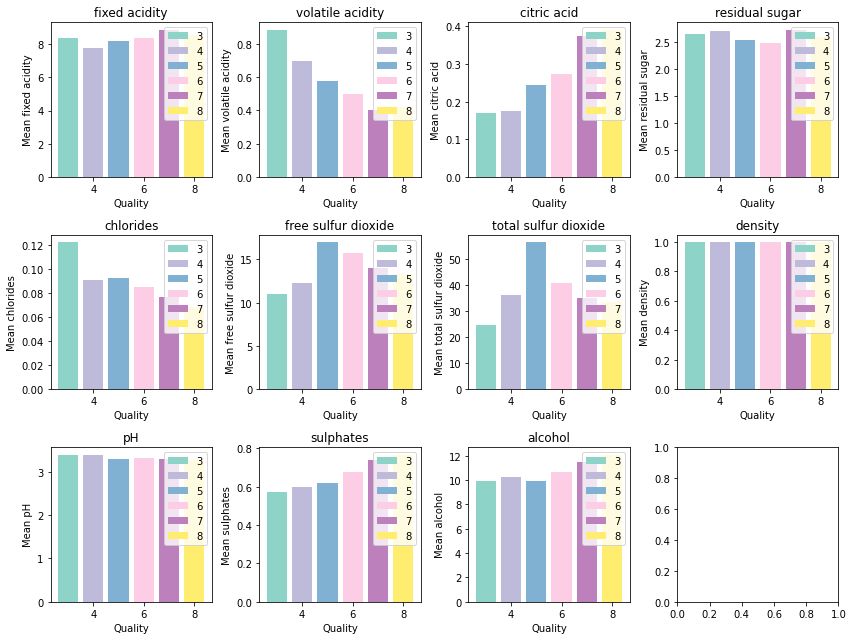
data, encoding data etc, all of this improves the data standard. It is not necessary to

utilize all of these preprocessing features. The dataset with 1599 samples is taken

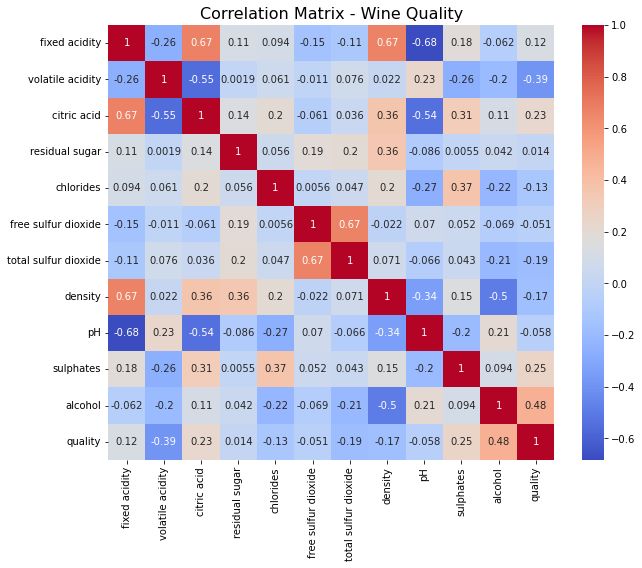
and randomly divided into 80% as the training set and 20% as the test data. Training data is used to best fit the model and test data is use to examine the performance of the model.

This is the process flow diagram model for wine quality prediction.

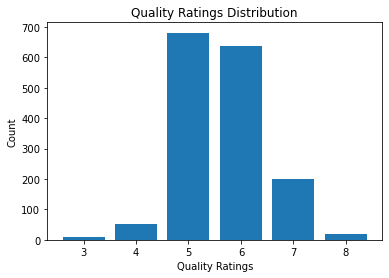




These subplots were generated,in which each subplot represents the mean values of a different feature grouped by the quality ratings in the 'wine' DataFrame. It provides a visual comparison of the mean values across different quality ratings for each feature.



A heatmap correlation matrix gives the relation between different features. The features having negative values are less correlated to each other. The cells having dark blue color are weakly correlated, and cells having red color are strongly correlated. Machine Learning Models which we have used in this model are Logistic Regression Algorithm (LR), Decision Tree (DT), Random Forest Algorithm (RF), KNN Classifier (KNN), Support Vector Machine (SVM) and Naive Bayes Classification (NB). The Evaluation Metrics used are accuracy , precision , recall.



The quality ratings range from 3 to 8, with 3 being the lowest and 9 being the highest.

The distribution is not perfectly balanced, with a higher concentration of wines receiving ratings in the range of 5, 6, and 7.

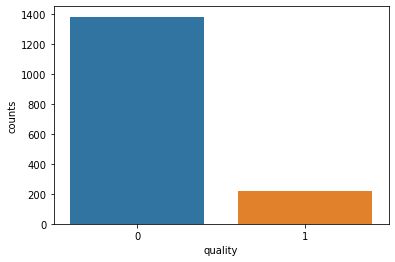
The lowest quality rating, 3, has the least number of occurrences, indicating that wines with this rating are relatively rare in the dataset.

The highest quality rating, 8, also has a relatively low occurrence, suggesting that wines with exceptional quality are less common.

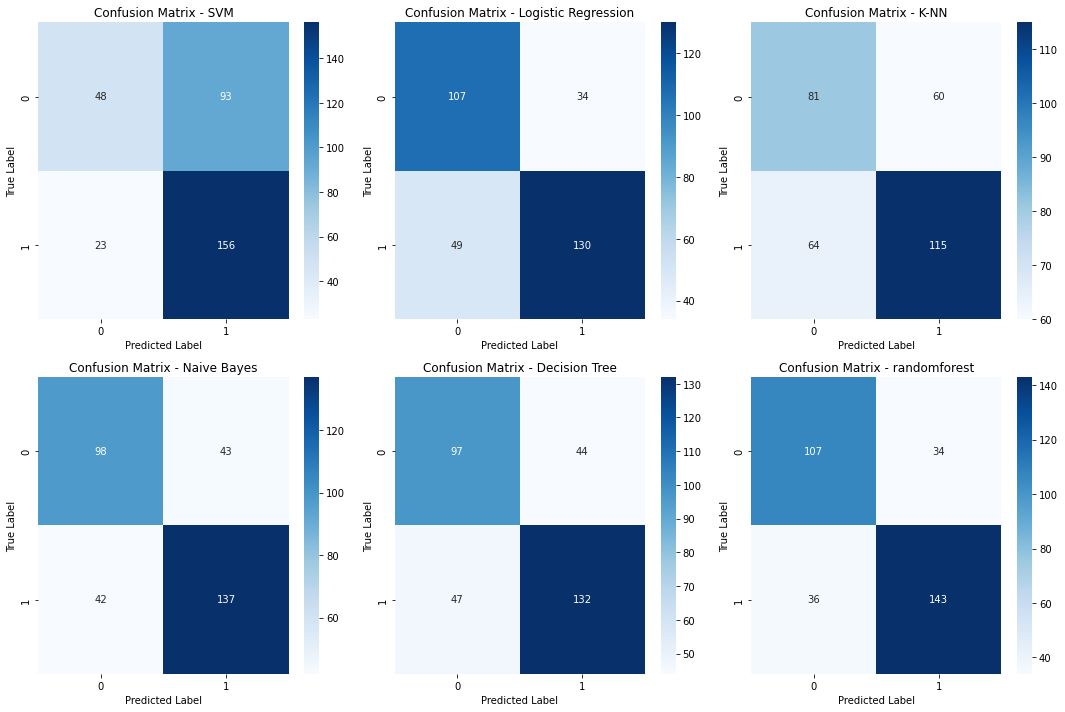
Inference:

Based on the quality ratings distribution, we can infer that the majority of wines in the dataset receive ratings between 5 and 7, indicating that they are considered average to above-average in quality. Wines with ratings below 5 or above 7 are relatively less common, suggesting that they represent either lower or higher quality outliers within the dataset.

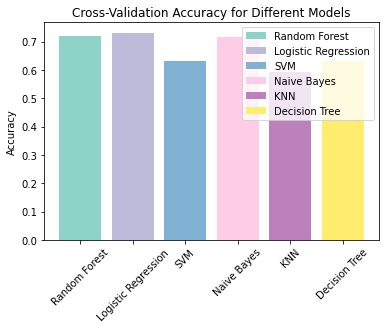
**3)DATA PROCESSING:-**



This plot has been generated by binning on the 'quality' column of the 'wine' Data Frame . The values of the quality of the wine between 2 and 6.5 will be labeled as 'bad', and values between 6.5 and 8 will be labeled as 'good'. By using Label Encoder object, which is used to convert the bin labels ('bad' and 'good') into numeric values. This step is necessary for certain machine learning algorithms that require numeric labels. This graph provides an overview of how many instances fall into each quality rating category. The bar plot provides insights into the distribution of instances across the different quality categories



This graph was generated by performing the binary classification by converting the target variable into binary labels and it can be viewed by trains multiple classifiers, predictions on the test set was generated, and the confusion matrices was calculated and visualized with the help of heatmaps. The confusion matrices provide insights into the performance of each classifier in terms of correctly and incorrectly classified instances.



Cross-validation helps to obtain a more reliable estimate of a model's performance. By considering multiple subsets of the dataset, cross-validation reduces the bias and variance that may be present in a single train-test split, providing a more robust evaluation of the model's performance.

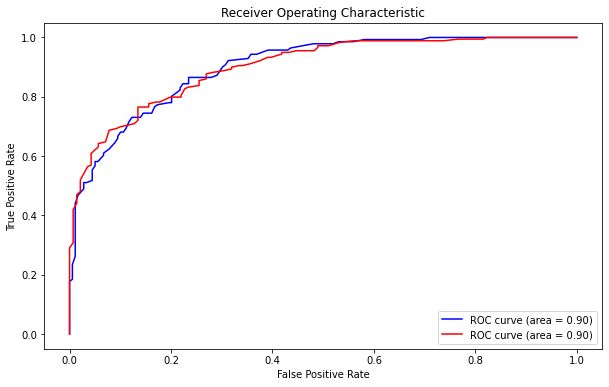
­­­

The bar plot of the cross-validation for multiple classification models was plotted with the help of accuracy scores.. This allows for a comparison of the performance of different models in terms of their cross-validated accuracy.

From this plot we can infer that the algorithms, LR, RF, NB, DT ,SVM and NB give us the accuracy of approximated to 73.1%, 72.1%, 71.7%, 63.2%, 63% and 59.2%.

Cross-validation helps to obtain a more reliable estimate of a model's performance. By considering multiple subsets of the dataset, cross-validation reduces the bias and variance that may be present in a single train-test split, providing a more robust evaluation of the model's performance.

From this cross-validation result we can infer that Logistic Regression is highly accurate for this data. And also, we have the Random Forest classifier slightly equal to the LR.



The ROC (Receiver Operating Characteristic) curve is a graphical representation used in machine learning to evaluate the performance of a classification model The ROC curve provides a visual representation of the performance of the RandomForestClassifier in distinguishing between different classes.

The AUC values quantify the classifier's performance, with higher values indicating better discrimination. A higher precision and recall indicate better performance for the classification model

From the ROC curve plot, you can make the following inferences:

The curve represents the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity).

The closer the curve is to the top-left corner, the better the classifier's performance.

The area under the curve (AUC) is a measure of the classifier's overall performance. A higher AUC indicates better discrimination between the classes.

Each color corresponds to a different class, and the AUC value is displayed in the legend for each class.

By comparing the AUC values for different classes, you can determine which classes are better separated by the classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **CLASSIFIERS** | **ACCURACY** | **PRECISION** | **RECALL** |
| Logistic Regression | 64.0625 % | 0.2992407732489 | 0.285964912280 |
| SVM | 67.1875 % | 0.32580103812980526 | 0.29844584286803966 |
| Decision Tree | 66.25 % | 0.3744912494912495 | 0.3469997457411645 |
| KNN | 60 % | 0.2804700328407225 | 0.27890922959572845 |
| Naïve Bayes Theorem | 56.875 % | 0.2875992877325937 | 0.30161454360539025 |
| Random Forest Theorem | 73.4375 % | 0.5421474556748529 | 0.3751461988304094 |

Thos tabulation gives the accuracy, precision, recall for various classifiers. Among the 6 classifiers Random forest theorem gave the highest accuracy. Finally ,considering the above results and observations, Random forest theorem reported the best performance and it is most suitable for this algorithmic study.