WINE QUALITY ANALYZER

IBM MINI PROJECT

Submitted by RITIKA M(9920004717)
V.SARANYA(9920004719)

P.V.GOWTHAMI(9920004635)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING COMPUTER SCIENCE AND ENGINEERING

KALASALINGAM ACADEMY OF RESEARCH AND

EDUCATION KRISHNANKOIL 626 126

Academic Year 2021-2022

CONTENTS

INTRODUCTION:

Wine consumption has increased rapidly over the last few years not only for recreational purposes but also because of its benefits to the human heart. Today, all industries are applying new techniques and implementing new methodologies to maximize production and making the whole process efficient. These processes are becoming expensive with time, and their demands are also increasing.Unlike wines have various purposes, but the chemicals used in them are more or less the same, but the type of chemicals used needs to be assessed, and hence, we adopt these methods to verify. Wine, once considered a luxury commodity, is today steadily liked by a extensive range of consumers. The 11th largest wine producer in the world is Portugal. Certification and evaluation of wine are essential elements in Portugal's wine industry which prevent contamination and are vital for quality assurance. Unlike old times, when there was a lack of resources and technology, the testing and quality assertion of wines couldn't be achieved, which is a critical aspect today because of the quality standards and to stay in the market is not easy, given the competition in the market. Wine has many attributes, such as pH, acidity, chlorides, sulphates, and other acids.

LITERATURE SURVEY:

Well, for research purposes, this could be described as the process of extracting secret information from the loads of databases. Information Discovery in Databases (KDD) is also known as Data Mining. Data Mining is commonly used in a variety of applications, like e-commerce, stock, product analysis, including understanding customer research marketing and real estate investment pattern etc. Data Mining is based on the mathematical algorithm required to drive the preferred results from the enormous collection of databases. Business Intelligence (BI) can be used for the analysis of pricing, market research, economic indicators, behavior use, industry research, geographic information analysis, and so on. Data mining technologies are commonly used in the fields of Customer Relationship Management, direct marketing, healthcare, e-commerce, telecommunications, and finance. This could also be likely you need to contact outsourcing companies for help. Such outsourcing firms are experienced in processing or scraping the data, filtering it out, and then keeping it for examination. Usually, data mining involves collecting information and analyzing the data and to search for more details etc.

A. Literature Survey Findings

1. Practically there is no impact on quality appears on the fixed acidity.

2. There are some negative connection with the quality which appears in volatile acidity.

3. There are many better wines available which appear to have higher grouping of Citric Acid.

4. There were some comparison is made in order to identify the better wines. These better wines appear to have higher liquor rates. Yet, when we made a direct model around it, from the R squared worth that liquor without anyone else just contributes like 20% on the difference of the quality. So there might be some different elements impacting everything here.

5. Even however it's a frail association, yet lower percent of Chloride appears to create better quality wines.

6. Better wines appear to have lower densities. In any case, of course, this might be because of the higher liquor content in them.

7. Better wines appear to be more acidic.

8. Residual sugar nearly has no impact on the wine quality.

.

DATA SET:

The first step before designing a model is data pre-processing. This step analyses the data distribution against various column. Null entry may be filled using mean and median of the particular column. Outlier detection is important to ensure the proposed classifier do not deviate from its prediction. The wine dataset considered in this chapter has no outliers, missing values, so it does not require intervention. However, a few rows repeated at the end. Such rows are done away with. There are two datasets, and both concerned with red and white samples of Vinho Verde wines from northern Portugal**.**

Features :

wine type - 1096 Red and 3451 White wine

fixed acidity - Most acids involved with wine or fixed or nonvolatile

volatile acidity - The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

citric acid - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

residual sugar - The amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

chlorides - The amount of salt in the wine

free sulfur dioxide - The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide - Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

density - the density of water is close to that of water depending on the percent alcohol and sugar content

pH - Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

sulphates - a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

alcohol - the percent alcohol content of the wine

Outcome Variable:

quality - score between 0 and 10

Feature Selection

In this section various feature variables in red white wine dataset are analyzed to detect the prevalent features to predict or assess the quality. The analyzed data in figure 2 and 3 offers a basis to derive the features importance for the analyzed model. Correlation tool is used select the dominant features based on the following facts. The variable fastened Acidity appears to possess virtually no impact on quality, Volatile Acidity shares a correlation with quality. The concentration of acid contributes the higher wine quality that direct correlation. An honest range against alcohol expected for quality wine. The lower concentration of Chloride appears to provide higher quality wines. Better wines used to be more acidic. And the residual sugar almost has no effect on wine quality. By analyzing the sample distribution in red and wine with nature of its independent variables the following feature are preferred to design the model. The important attributes are as follows acidity, sugar content,chlorides, sulfur, alcohol, pH and density.

CLASSIFICATION:

Cataloguing is an information mining highlight that relegates objects to target classifications or classes inside a set. The arrangement objective is to anticipate the objective class precisely in the information for every function. A grouping model might be utilized, for instance, to order advance candidates as little, medium, or high credit chances. Arrangement errands start with an informational collection that knows the class tasks. Characterization is discrete and doesn't infer request. Nonstop, skimming **point**

esteems will suggest an objective number rather than a clear cut one. A prescient model that has a mathematical objective uses a relapse calculation, not a calculation for order. The clearest kind of issue with order is a double grouping. The objective quality in paired characterization has just two potential qualities: high praise score or low praise assessment, for instance. Multiclass targets have multiple qualities: low, medium, high, and obscure FICO ratings, for instance. In the model build strategy (preparing), an arrangement calculation discovers connections between the indicator esteems and the objective qualities. Various calculations for the arrangement utilize explicit strategies to recognize connections. These connections are plot in a model that would then be able to be applied to another arrangement of information in which the class tasks are obscure. Characterization models are assessed by contrasting the normal qualities in a bunch of test information against realized objective qualities. Scoring a measure of classification results in in-class assignments and probabilities for each particular event. For example, the likelihood of each classification for each customer can also be predicted by a model which classifies customers as low, medium, or high.

Consequently, the goal of the proposed chapter is to predict the quality of the wine based on physicochemical tests through machine learning models. The upcoming sections precisely narrate the classification steps adopted by them inprediction.

DECISION TREE:

A Decision Tree is a typical portrayal, for instance, arrangement. It is a Supervised Machine Learning where the information is constantly isolated by a particular boundary. A progression of preparing models is separated into more modest and more modest subsets while simultaneously steadily making a connected choice tree. A choice tree that covers the preparation set is returned toward the finish of the learning cycle. The key thought is to utilize a choice tree to segment the information space into (or thick) group areas and vacant (or scanty) districts. Another model is ordered in Decision Tree Classification, by sending it to a progression of tests that choose the model's class mark. In a various levelled framework called a choice tree, such tests are requested.

A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality317Published By:Blue Eyes Intelligence Engineeringand Sciences Publication Retrieval Number: 100.1/ijrte.A58540510121 DOI: 10.35940/ijrte.A5854.0510121 Choice Trees obey Algorithm of Divide-and-Conquer. Decision trees are manufactured utilizing heuristic parcelling, called recursive apportioning. This strategy is additionally for the most part alluded to as separating and vanquishing since it partitions the information into subsets, which are then more than once isolated into much more modest sub-sets etc., until the cycle stops when the calculation concludes that the information in the sub-sets are adequately homogeneous or has another halting measure. Utilizing the choice calculation, we start from the base of the tree and split the information on the element that outcomes in the main increase of

data (IG) (decrease of vulnerability towards the decision).Then we can rehash this parting strategy at every youngster hub in an iterative cycle until the leaves are unadulterated, which implies the examples at every hub of the leaf are the entirety of a similar class. By and by, to forestall overfitting, we can set a cut-off on the tree's profundity. Here we rather bargain on virtue, as the last leaves can in any case have some pollution.

RANDOM FOREST:

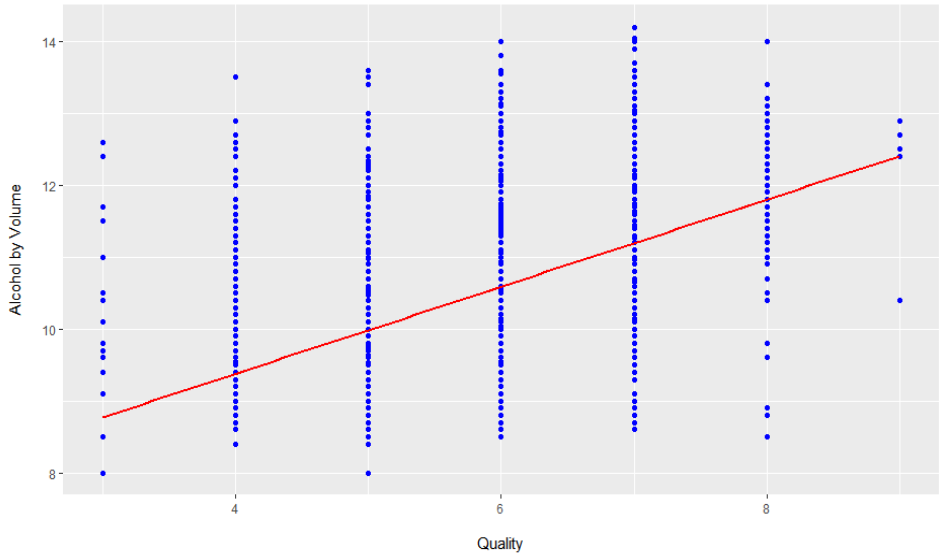
Random Forest Algorithm is an administered algorithmfor the grouping. We can see it from its name, which somehow or another, is to make a forest and make it arbitrary. There is an away from between the quantity of trees in the forest and the outcomes it can acquire: the more noteworthy the quantity of trees, the better the outcome. In any case, one thing to recall is that building the forest isn't equivalent to settling on the choice with information increase or record strategy.

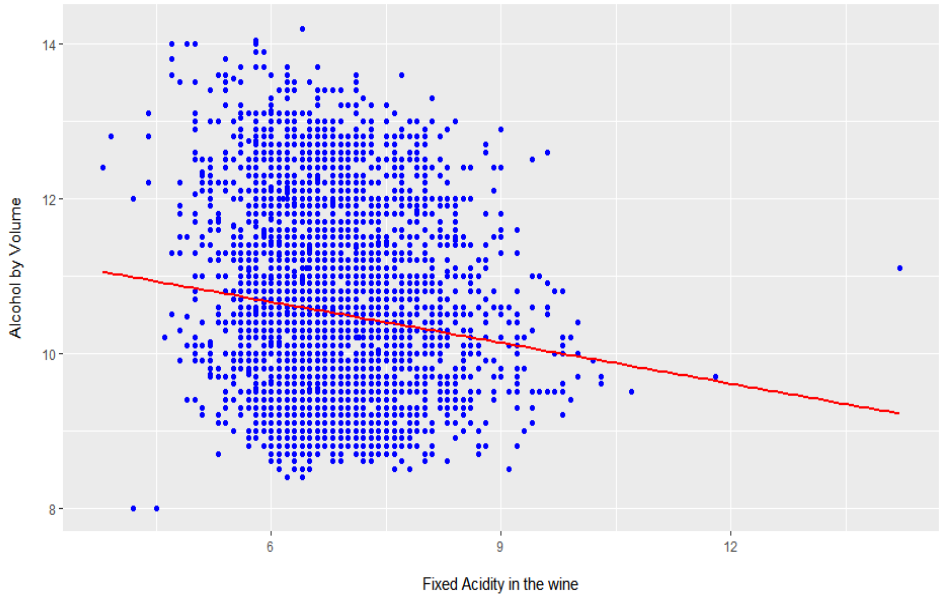
The contrast between the Random Forest calculation and the choice tree calculation is that the way toward finding the root hub and isolating the component hubs would run haphazardly in Random Forest. It tends to be utilized for assignments identified with grouping and relapse. Overfitting is one significant issue that can exacerbate the outcomes, yet for the Random Forest calculation, the classifier won't over fit the model if there are sufficient trees in the timberlands. The third favorable position is that the Random Forest classifier can oblige missing qualities, and the last bit of leeway is that unmitigated qualities can be displayed on the Random Forest classifier. In the Random Forest algorithm, there are two phases, and one is random forest creation, the other is to make an expectation from the irregular random forest classifier produced in the main stage. The whole cycle is appeared beneath and utilizing the figure and it is easy to comprehend.

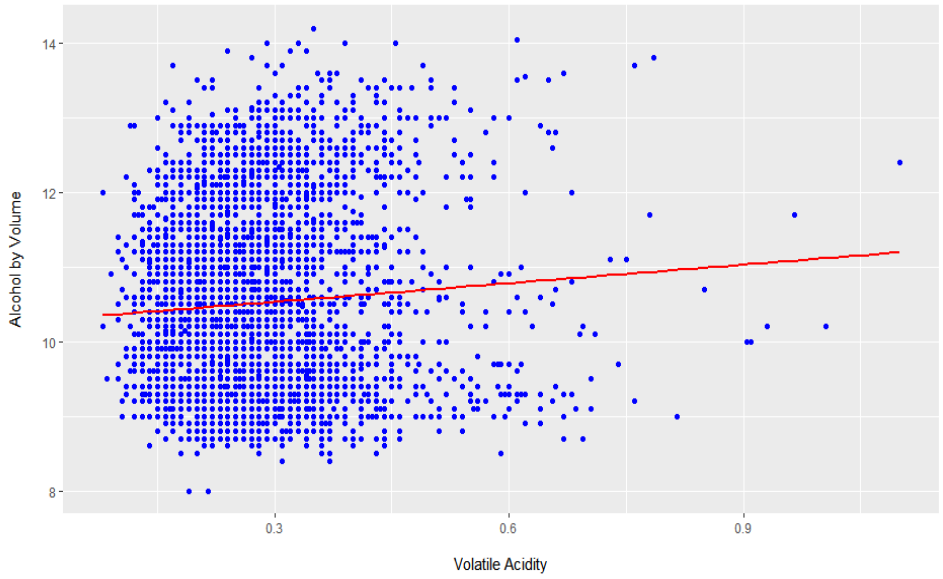
Pseudo Code - Random Forest
• Randomly pick "m" highlights from all out "n" highlights where m< < n ○Among the "m" highlights, measure the"d" hub utilizing the best part point ○Break the hub into youngster hubs utilizing the best split. • Repeat the previously mentioned ventures until the predetermined number of hubs are reached.

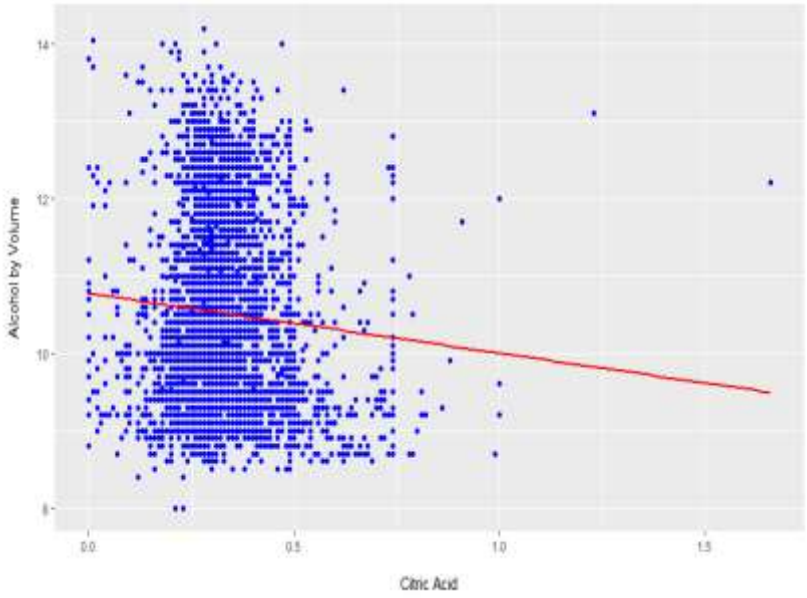
- Creating timberland by rehashing steps previously mentioned steps to construct "s" number of trees.

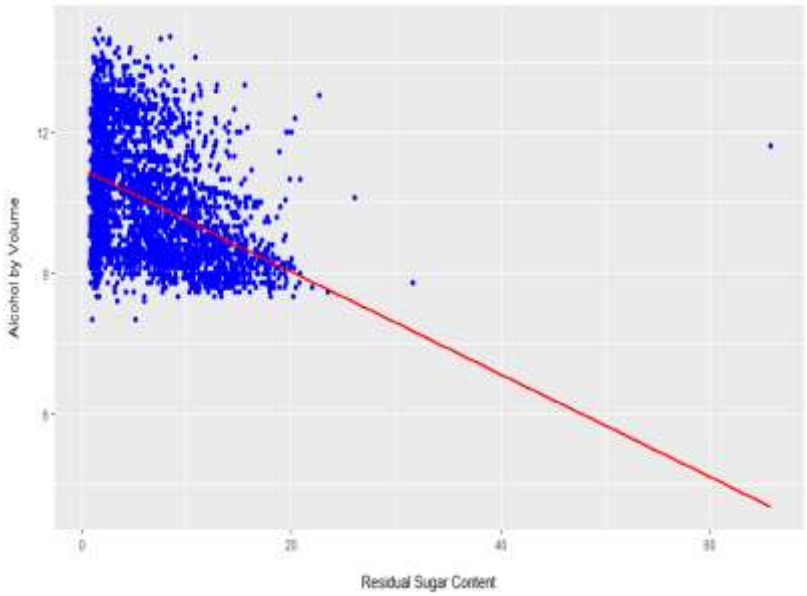
CORRELATION BETWEEN ALCOHOL AND OTHER VARIABLES:

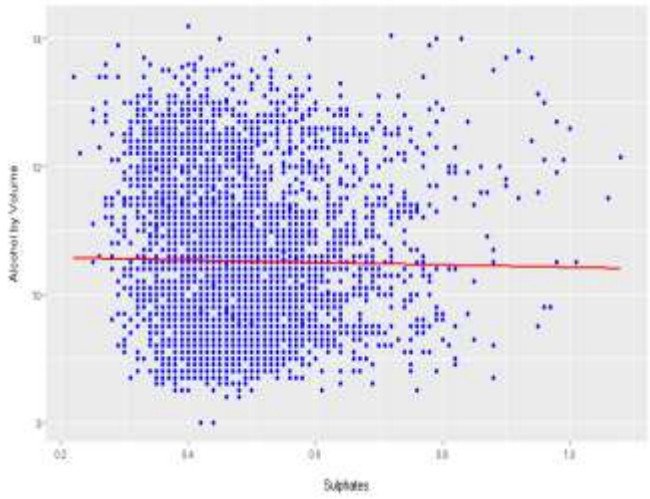



Alcohol by Volume
14
12
10
8
6
9
12
Fixed Acidity in the wine


Alcohol by Volume
14
12
10
8
0.3
0.6
0.9
Volatile Acidity

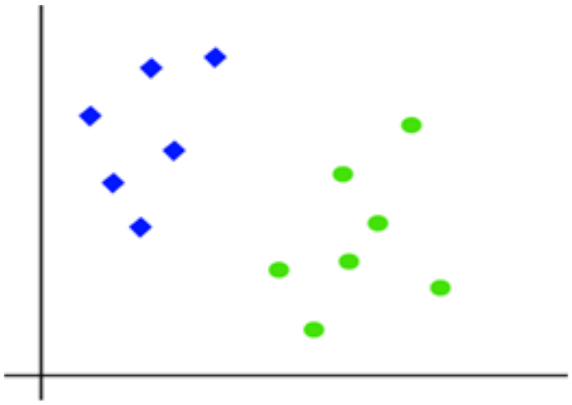

Alcohol by Volume
14
12
10
8
0.0
0.5
1.0
1.5
Citric Acid


Alcohol by Volume
12
8
0
20
40
60
Residual Sugar Content



SUPPORT VECTOR MACHINE:

Support Vector Machine is one of the most well-known Supervised Learning algorithm utilized for grouping just as for issues with relapse. In Machine Learning, be that as it may, it is utilized principally for arrangement issues. The SVM's algorithm will likely form the best line or choice limitthat can isolate n-dimensional space into classes so that lateron, we can advantageously situate the new information pointin the proper classification. This best limit for judgment is known as a hyperplane. SVM chooses the outrageous focuses/vectors which help to build a hyperplane. Theseoutrageous cases are alluded to as help vectors, and subsequently the calculation is called Support VectorMachine. Assume we have a dataset with two labels (green and blue), and two x1 and x2 capacities in the dataset. We need a classifier which groups the directions pair(x1, x2) intoeither green or blue. Consider the underneath picture in Figure 3



TWO CATEGORY CLASSSIFICATION.

So as this is 2-d space, we can easily distinguish these twogroups by simply using a straight line. Yet those classes can be separated by several lines. Consider the below image in Figure 4. Hence, the SVM algorithm helps find the best line or boundary of decision; this best boundary or region is calleda hyperplane. SVM algorithm from both groups seeks the closest point of the row, such points are called vectors of support.

The distance between the hyperplane and the vectors is called margin. SVM's goal is to optimize the margin depictedin Figure 5.

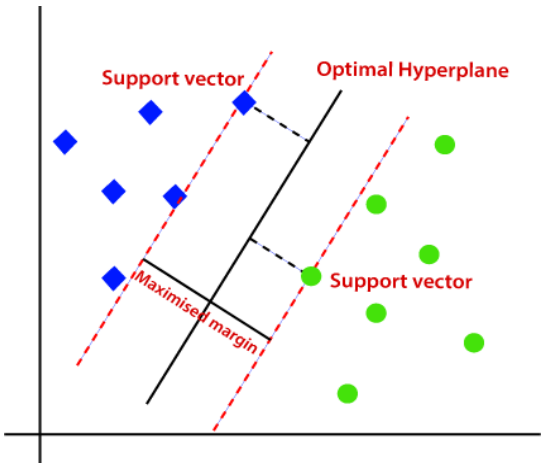


Figure 5. SVM defines boundary using hyperplanE.

Pseudo code:

The steps are used to implement the wine quality prediction model is depicted as pseudo code

Pseudo code: Wine quality rate computing system

Input: Red and White wine datasetOutput: Quality score

Step 1: 3.Load the datasets

Step 2: Summarize the data distribution range using Visualization tool 5.Creating a variable "rating" for white andred wine respectively

Step 3: Identify the prevalent features using correlation tool**Step** 4: Split the input dataset into train and test using 75:25ratio

Step 5: Transform the data to fed into machine learningmodels

Step 6: Invoke Decision-Tree() **Step** 7: Invoke Random Forest()**Step** 8: Invoke KNN ()

Step 9: Invoke SVM()**Step** 10: Invoke LR() **Step** 11: Invoke MP5

Step 12: Summarize the performance in terms rating andstrength of a models using metrics.

Evaluation Metrics

The following are the metrics followed for evaluating thequality of the machine learning algorithms.

Confusion Matrix

A confusion matrix is a bench that is frequently cast-off to **label the presentation of a classification model** (or "classifier") on a usual of examination data for which the correct standards are recognized.

Where, x=y àD =0, x≠yà=1

(2)

Confidence Interval (CI)

A confidence intermission, in data, mentions to the likelihood that a residents parameter will drop among two fixed standards for a sure amount of eras. Confidence

The easiest way to select the precise value for K is by evaluating the data first. A high K worth is usually additionalreliable as it decreases the total noise but here is no assurance.Cross-validation is additional method to evaluate a successfulK worth retrospectively, by an self-governing dataset to test the K value. Factually the ideal K has remained between 3-10for most datasets. The results are better than 1NN.

Balanced accuracy

It is metric that unique container use when assessing in what way decent a two classifier is. It is particularly valuablewhen the lessons are unfair, i.e. unique of the binary cases seems a ration additional frequently than the additional.

Results

The analysed wine quality rate computing system is simulated in R programming environment. Quality is anattribute which defines the quality as rated by the wineexperts Fig 6. It is an integer between 0 to 10, 0 being the lowest and 10 being the highest. As we can see in the graphs,

the maximum distribution is between 5 and 6, which is the average of the quality index thus we can infer that the majority of the wines present in dataset is average with very low good and worst quality wines. To achieve better results, we define a variable called rating from quality where if the quality is less than 5, the rating is classified as bad and if lessthan 7, then as average or good and above 7 as good. Therefore, we can conclude from the graph that the majorityof the wines which we have are of average quality andreliable for tests.

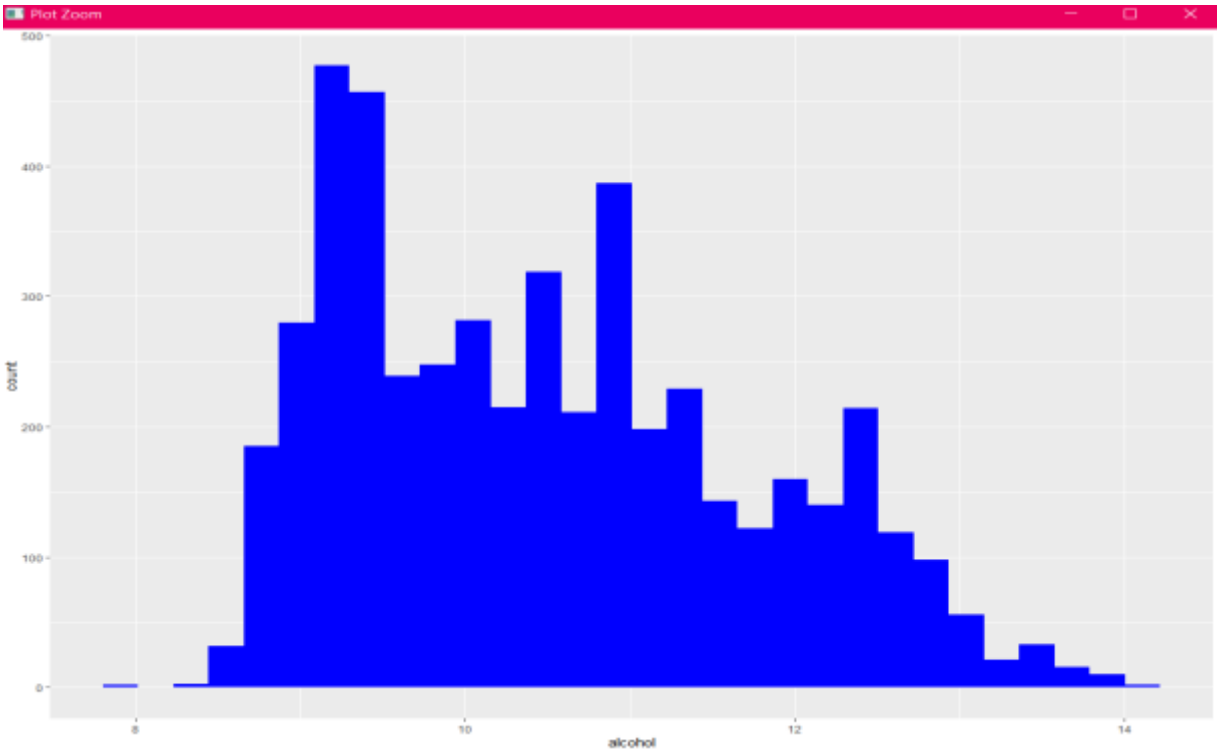
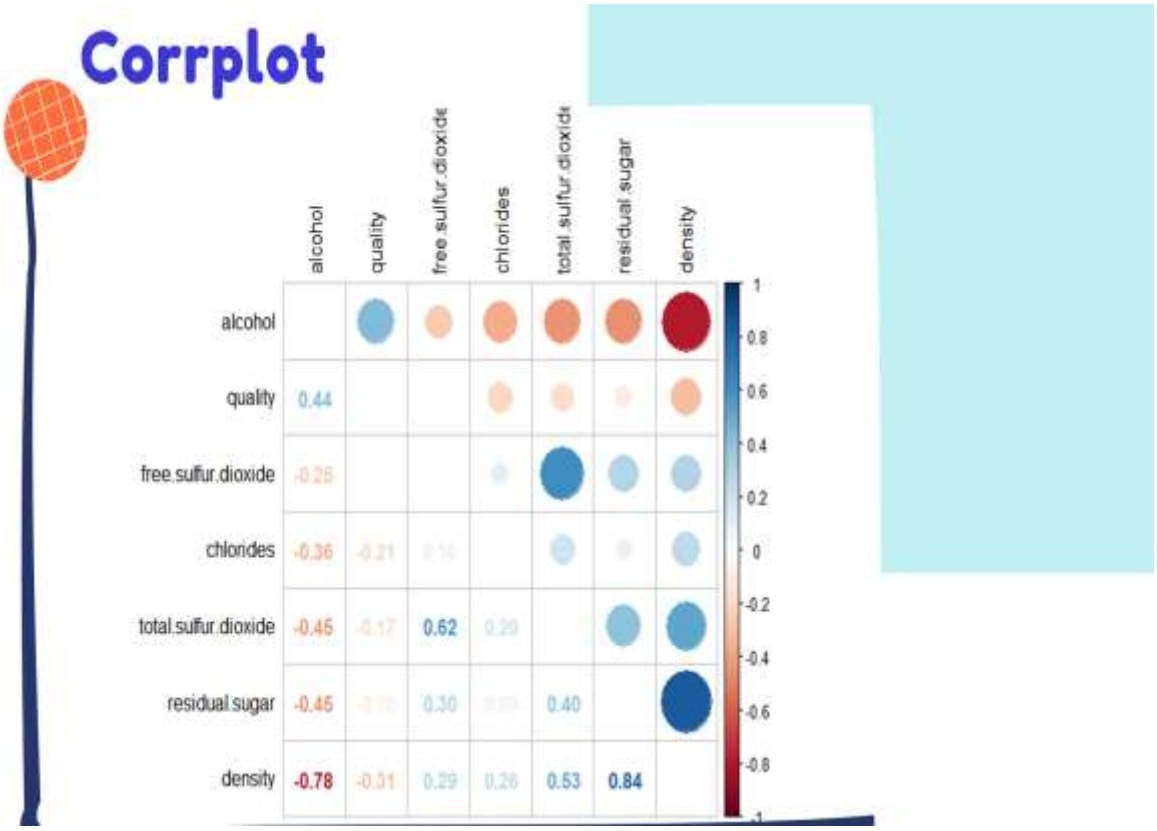


Figure 6 Qualities of Wine Ratings by Experts



CONCLUSION:

We suggest that wine producers should concentrate on maintaining an acceptable alcohol content through longer fermentation times or higher fermentation yeast yields. In recent years, interest has increased in the wine industry, which is demanding growth in this market. The companies are also investing in new technology to increase wine quality and sales. Wine quality certification plays a significant role in all processes in this direction, and it needs human experts to test wine. This paper goes through the use of two classification algorithms, Decision Tree and Random Forest algorithms are implemented on the dataset, and the two algorithms' output is compared. Results showed that our improved MP5 (Multiple Regression Model) had outperformed the Decision Tree and Random Forest techniques, particularly in the most common type of red wine. There are two sections of this dataset: Red Wine and White Wine. There are 1599 samples of red wine and 4898 samples of white wine. The dataset of both red and white wine is composed of 11 physicochemical properties. This work deduces that the classification method should provide space for corrective steps to be taken during

production to enhance the quality of the wine. In the future, broad data set may be used for experiments and other machine learning techniques may be explored for prediction of wine quality, and we will expand this analysis to include feature development methods to test whether or not the model's predictive power may be increased.

STATUS OF THE PROJECT:

CLASSIFIACTION AND QUALITY ANALYSIS – COMPLETED

IMPLEMENTATION OF CODE – COMPLETED

REFERENCES:

1. Gupta, Y., 2018. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, pp.305-312.

2. Sun, Danzer and Thiel. (1997) "Classification of wine samples by means of artificial neural networks and discrimination analytical methods". Fresenius Journal of Analytical Chemistry359 (2)143–149.

3. Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009, October. Using data mining for wine quality assessment. In International Conference on Discovery Science (pp. 66-79). Springer,Berlin, Heidelberg.

4. Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) "Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks". Talanta 72 263–268.

5. Yu, Lin, Xu, Ying, Li and Pan. (2008) "Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy".

6. Agricultural and Food Chemistry 56 307–313. [7] Beltran, Duarte-Mermoud, Soto Vicencio, Salah and Bustos. (2008) "Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer". IEEE Transactions on Instrumentation and Measurement 57 2421-2436.

7. Cortez, Cerdeira, Almeida, Matos and Reis. (2009) "Modeling wine preferences by data mining from physicochemical properties". Decision Support Systems 47 547-553.

8. Jambhulkar and Baporikar. (2015) "Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network". International Journal of Computer Science and Applications 8 (1) 55-59.

9. Zaveri, and Joshi. (2017) "Comparative Study of Data Analysis Techniques in the domain of medicative care for Disease Predication". International Journal of Advanced Research in Computer Science 8 (3) 564-566.