A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality

Article *in* International Journal of Recent Technology and Engineering (IJRTE) · May 2021

DOI: 10.35940/ijrte.A5854.0510121

2 authors, including:

Vanmathi Chandrasekaran
VIT University
49 PUBLICATIONS **140** CITATIONS

SEE PROFILE

A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality

Mohit Gupta, Vanmathi C

Abstract: ***In today's trend consumers are very much concern about the quality of the product in turn, Industries are all working on various methodologies to ensure the high quality in their products. Most of consumers judge the quality of the product based on the certification obtained for the product. In Earlier days, the quality is measured and validated only through human experts. Nowadays most of the validation tasks are automated through software and this ease the burden of human experts by assisting with them in predicting the quality of the product and that leads to greater a reduction of time spent. Wine consumption has increased rapidly over the last few decades, not only for recreational purposes but also due of its inherent health benefits especially to human heart. This chapter demonstrates the usage of various machine learning techniques in predicting the quality of wine and results are validated through various quantitative metrics. Moreover the contribution of various independent variables facilitating the final outcome is precisely portrayed.***

Keywords: Machine Learning, KNN, Random Forest, SVM, J48, Wine Quality.

I. 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. Wine quality can be measured either using physiochemical test, or sensory test. The former test can be established without the human intervention, whereas the later can be possible under the supervision on human expert. Manufacturers will use the predictions from this model to improve wine quality, certification agencies to better understand the factors that are essential for quality and to allow consumers to decide while purchasing it. There many authors designed several predictive models to assess the wine quality as part of automation. [1] used linear regression to detect the predominant features. These set of features fed into as input support vector machine and neural network models values of output variable. The proposed techniques experimented on White Wine and Red Wine datasets. Authors concludes the better prediction heavily rely on the correct set of feature variables. The three-layer Back propagation neural network (BPN) model is proposed model to categorize wine samples from six various geographical areas which is established on the basis of measurements of independent variables. The prediction quality of the proposed model is improved by optimizing the hyper parameters. The performance of the classifier is proved by comparing the results with bench mark methods [2]. Three regression techniques were applied. The model selection and variable selection is done through sensitivity analysis. The results obtained through support vector machine outperforming then the neural network methods and multiple regression. The proposed model is beneficial for testing the influences of the sensory preferences[3].

Moreno et. al [4] categorized 54 samples of wine using a probabilistic neural network into two levels. Authors Yu et al. [5] have listed 147 bottles of rice wine using spectral measurements to estimate three types of wine. For the classification of Chilean wine, Beltran et al. [6] used SVM, linear discriminate analysis and neural network. The studies and analyzes were carried out on three different Chilean wine varieties. Cortez et al. [7] has related various dataset classifications in wine.

Jambhulkar et al. [8] used several strategies to forecast heart disease using a network of wireless sensors. They extracted the attributes for heart disease prediction from the Cleveland dataset. Zaveri et al. [9] used data-mining techniques to predict various diseases such as tuberculosis, cancer, diabetes, etc. Beltrán et al. [10] proposed a wine classification method focused on aroma chromatograms to extract features and classifiers such as neural network, linear discriminating analysis, and supporting vector machine and found that supporting wavelet transforming vector machine performs better.

Manuscript received on May 12, 2021.
Revised Manuscript received on May 31, 2021.
Manuscript published on May 30, 2021.
* Correspondence Author
Mohit Gupta*, School of Information Technology and Engineering Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: mgpta007@gmail.com
Vanmathi C, School of Information Technology and Engineering Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: vanmathi.c@vit.ac.in

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Thakkar et al. [11 ranked the attributes using analytical hierarchy and then used a different classification of machine learning such as supporting vector machine , random forest and SVM. Reddy et al [12] recommended a user-centric clustering approach and collected the red wine data for survey purposes.

In literature, various researchers have used different methods and technologies to maximize accuracy, but much scope of improvement yet to be achieved. This paper stresses effort on many aspects; this phase consists of data collection and preliminary data analysis to assess the quality of the data. It includes cleaning up the data, making other transformations on the data in order to achieve the final data collection. Modeling involves choosing the appropriate modelling technique. Evaluation models produced during the modelling process are evaluated for quality. Deployment models are finally put to use during this process.

II. 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 software solution could be the best choice for data mining as it will save an enormous amount of time and effort. Free Text Software Technologies, LexiQuest, Megaputer Text Analyst, Connexor Machines, SAS Text Miner, Lextek Profiling Engine, WordStat, are some of the popular Data Mining software programs available.

Data mining helps to forecasts accurate and reliable historical data that we have at our fingertips and guessing about future outcomes. Businesses may use data mining to assess why it is necessary to find the information they need to use business intelligence to analytics. They considered the gut micro biome of red wine drinkers to be more diverse than the non-red wine drinkers. He was not found with the consumption of white wine, beer, or spirits.

Caroline et al explains that moderate consumption of red wine is related with greater diversity and explains the effects on health. The microbiome is a group of microorganisms and plays a significant role in human health. The gut microbiome of an individual with a greater number of different bacterial species is considered a marker of gut health. The team observed that there was a greater number of different bacterial species in the gut micro biota of red wine consumers compared to non-consumers. This finding was also found in the UK, the USA, and Belgium in three separate cohorts. The authors took into account variables like age, weight, daily diet, and the participants' socioeconomic status and proceeded to see the association. The authors conclude that due to the many polyphenols in red wine, the principal explanation for the connection is. Polyphenols are naturally found in many fruits and vegetables as defensive chemicals.

Tim et. al [12] discussed the effects of red wine on the guts of nearly three thousand people in three different countries , they found that polyphenols in grape skin could be responsible for much of the controversial health benefits when used in moderation." The study also discovered that lower levels of obesity and 'bad' cholesterol were associated with red wine consumption, partly due to gut micro biota. "Even though we have established a correlation between the consumption of red wine and the diversity of gut micro biota, drinking red wine rarely, like once every two weeks, seems sufficient to detect an impact. Though, alcohol consumption with moderation is still advisable, “Dr. Le Roy added.

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.

III. PROPOSED FRAMEWORK

The proposed framework Figure 1 demonstrated in this chapter is discussed in this section.

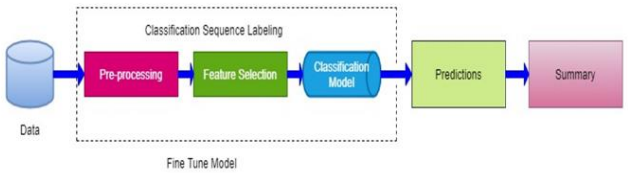


Figure 1. Proposed Framework

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

3.1 Dataset

The dataset [12] has 12 features. Among twelve features 11 features are independent variables and one dependent variable. Two categories of wine analyzed red wine and white one. In our analysis we have considered nearly 1000 and above samples, among that 1599 samples were red wine and 4898 samples were white wine. Figure 2 shows the feature variable chlorides distribution found in red and wine datasets. The description of feature variables can be found in Table 1.

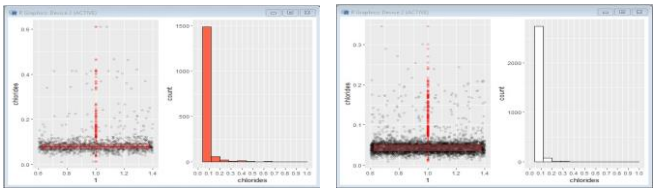


Figure 2. Data distribution statistics of various independent variables in wine dataset.

3.2 Data Pre-Processing

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.

3.3 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.

Table 1. Independent and Dependant variables description for Wine dataset

Features	Description	Features	Description
FA1	Fixed Acidity	FA7	Total sulphur dioxide
FA2	Volatile Acidity	FA8	Density
FA3	Citric acid	FA9	PH
FA4	Residual sugar	FA10	Sulphates
FA5	Chlorides	FA11	Alcohol
FA6	Free sulphur dioxide	OT1	Score

3.4 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 in prediction.

3.4.1 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.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

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.

3.4.2 Random Forest:

Random Forest Algorithm is an administered algorithm for 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.

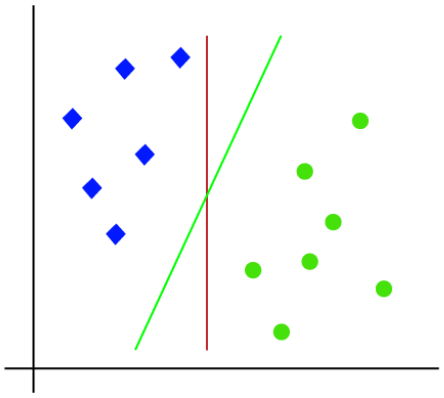
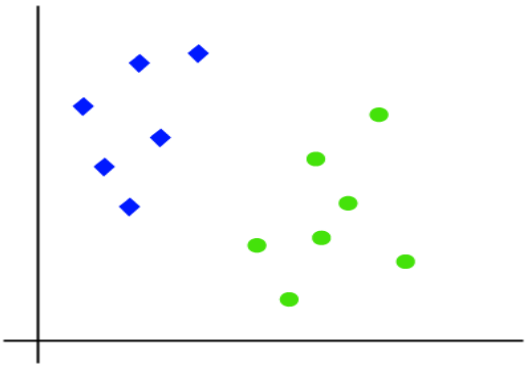


Figure 4. SVM defines boundary using hyperplane

3.4.3 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 limit that can isolate n-dimensional space into classes so that later on, we can advantageously situate the new information point in 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. These outrageous cases are alluded to as help vectors, and subsequently the calculation is called Support Vector Machine. 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) into either green or blue. Consider the underneath picture in Figure 3



Figurer 3. Two category classification

So as this is 2-d space, we can easily distinguish these two groups 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 called a 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 depicted in Figure 5.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

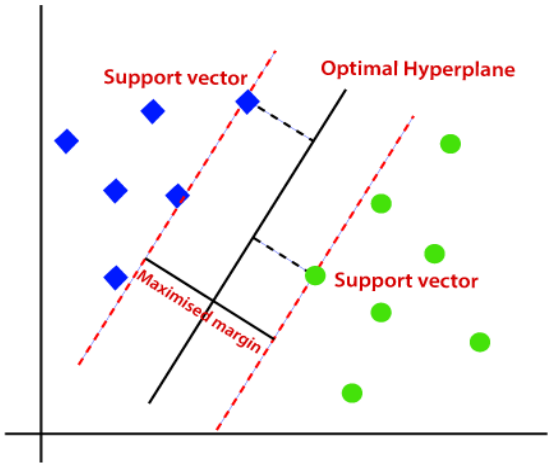


Figure 5. SVM defines boundary using hyperplane

3.4.4 KNN

K Nearest neighbors is a humble procedure that supplies all obtainable belongings and categorizes novel cases founded on a degree of resemblance (e.g., distance functions). KNN was previously used as a non-parametric method in arithmetical approximation and pattern recognition in the initial 1970s. A circumstance is classified by a widely held vote of its neighbours, the circumstance actuality allocated by a distance method as stated in equation (1) to the greatest mutual class of its adjacent K neighbour's. If K = 1, formerly the circumstance is assigned to its neighbouring neighbour's class.

$$DistFn = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (1)$$

It must also be recalled that altogether three events of distance are usable only for non-stop variables. The Hamming distance must be used for the example of definite variables. It also raises the problem of standardizing the arithmetical variables among 0 and 1 once a combination of arithmetical and definite variables is present in the dataset.

$$D_H = \sum_{i=1}^{k}|x_i - y_i| \quad (2)$$

Where, x=y →D =0, x≠y→=1

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

3.3.4 MP5 Model

Model tree MP5 is a combination of data classification and regression. It follows the idea of methods for the decision tree, but instead of class labels, it has liner regression functions in the leaves.

The MP5 model tree is designed in a top-down manner. A decision is taken at each stage whether to partition the training set (i.e., create a split node) or add a regression function as a leaf node. The decision is based on the target variable standard deviation.

3.3.5 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 dataset
Output: 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 and red wine respectively
Step 3: Identify the prevalent features using correlation tool
Step 4: Split the input dataset into train and test using 75:25 ratio
Step 5: Transform the data to fed into machine learning models
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 and strength of a models using metrics.

3.5 Evaluation Metrics

The following are the metrics followed for evaluating the quality 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.

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 intervals quantify the degree of doubt or cert in a sample method. A confidence interval can yield any amount of likelihoods, with the greatest shared being a 95% or 99% confidence level. A *Confidence interval* is a choice of standards that probable would cover an unknown populace parameter.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

Classes

The prediction system described in this chapter is a multi-class problem and attempt to classify the wine quality into one of the three rating categories: 1-Good (quality 7 or above), 2-Average (quality of 5-6), and 3-Cheap (quality 4 or below). Therefore, classes 8 and above represent good quality wine, classes 5 to7 represent average wines and classes 3 and 4 show cheap quality wines.

No Information Rate

The no information error rate is the error rate when the input and output are independent. Accuracy should be higher than No information rate (naive classifier) in order to be model significant.

p-value

This degree is recycled to examination the correctness is improved than the "no info rate," which is occupied to be the main class fraction in the statistics.

The kappa coefficient events the contract among classification and fact values. A kappa value of 1 represents faultless contract, though a worth of 0 signifies no contract. McNemar's test is exactly a test of *paired* sizes.

Sensitivity

Recall is clear as the relation of the entire amount of properly categorized positive classes split by the total amount of optimistic lessons.

Specificity

Specificity controls the amount of actual rejections that are appropriately recognized

Pos and neg pred value

The optimistic prognostic worth is clear as the fraction of foretold positives that are really positive however the bad predictive worth is clear as the percentage of bad positives that are really bad.

Prevalence

How often does the yes condition actually occur in our sample? Actual yes/total.

Detection Rate

The probability of getting detected among all cases that is as represented below

Probability of detection prevalence = Sum of True positive/Sum of Condition Positive

Balanced accuracy

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

3.6 Results

The analysed wine quality rate computing system is simulated in R programming environment. Quality is an attribute which defines the quality as rated by the wine experts 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 less than 7, then as average or good and above 7 as good. Therefore, we can conclude from the graph that the majority of the wines which we have are of average quality and reliable for tests.

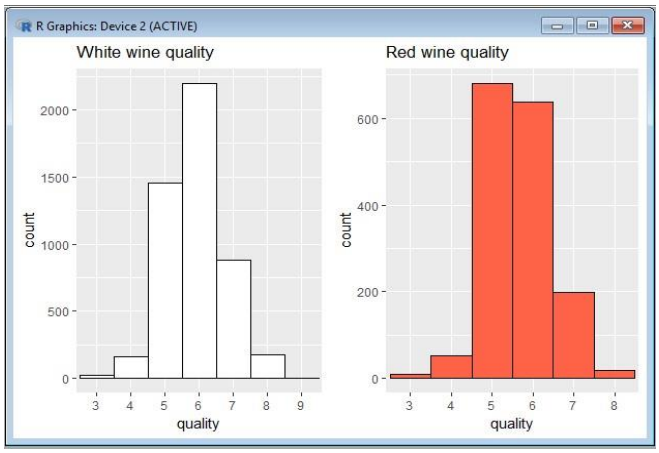


Figure 6 Qualities of Wine Ratings by Experts

3.5 Comparative Results

The below table 2 and 3 summarizes the performances of various machine learning algorithms.

Table 2 Quantitative performance comparison for Red wine dataset

Methods	Accuracy	Miscalculation Rate
J48	0.56	0.44
KNN	0.6139	0.3861
SVM	0.6234	0.3766
CART	0.7075	0.2925
Random Forest	0.7325	0.2675
M5P	0.8181	0.1819

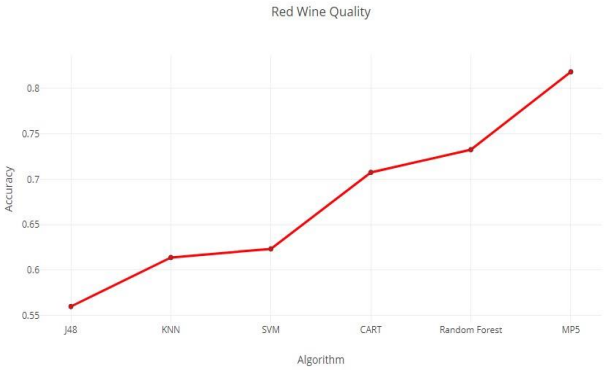


Figure7 Comparison Red Wine Dataset

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

While observing the results of various machine learning techniques using white wine samples only one variant, the Random Forest variant, performed better. Random Forest returned 68.4±2.3 accuracy and 49.9 kappa accuracy. K-nearest neighbors performed statistically worse, and the SVM model revealed that the 95 percent confidence interval was not substantially better or worse.

While in the case of samples in red wine, as with the data on white wine, the Random Forest model was the only one that did better than the benchmark. Accuracy of the Random Forest is 68.7±4.0, with a Kappa of 49.6. The K-nearest neighbors and accuracies of the SVM model were neither statistically better nor worse than the comparison. For both the datasets, MP5 model performed better. Since white wine is more prone to changes in physicochemical properties than red wine, we suggest a higher degree of disassociation between white wine and red wine production line with unusually more customization of white wine. The selection algorithm attribute we applied also rated alcohol as the highest in both datasets, making alcohol level the key attribute that defines both red and white wine quality.

Table 3 Quantitative performance comparison for White Wine dataset

Methods	Accuracy	Miscalculation rate
KNN	0.6052	0.3948
SVM	0.6372	0.3628
J48	0.6936	0.3064
Random Forest	0.7639	0.2361
CART	0.7819	0.2181
M5P	0.8327	0.1673

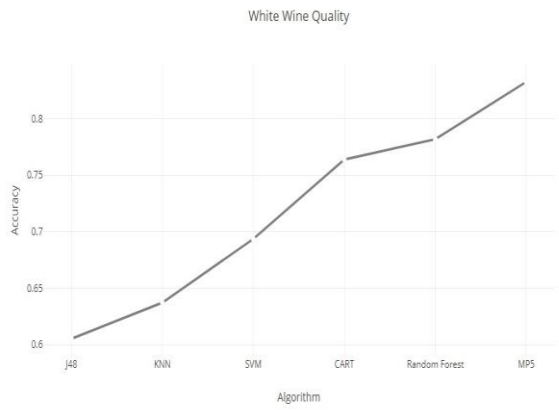


Figure 8 Comparison White Wine dataset

IV. CONCLUSION AND FUTURE SCOPE

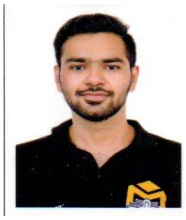
We suggest that wine producers should concentrate on maintaining an acceptable alcohol content through longer fermentation times or higher fermentation yeast yields. In recent years, interest has increased in the wine industry, which is demanding growth in this market. The companies are also investing 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.

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.
10. https://data-flair.training/blogs/r-decision-trees/
11. https://www.sciencedaily.com/releases/2019/08/190828194219.htm
12. https://www.kaggle.com/sgus1318/winedata

AUTHORS PROFILE



Mohit Gupta, A 2020 pass out from VIT University, Vellore with a bachelors in Information Technology. Currently working as a Business Analyst in Bewakoof.com with interest in business analytics and information management systems.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Dr. Vanmathi C, received her Ph.D. degree in Information Technology and Engineering from VIT University, M.Tech (IT) from Sathyabama University and B.E. Computer Science from Madras University. She is working as an Associate Professor in the School of Information Technology at VIT University, Vellore Campus, India. She is having 15 years of research experience. Her area of research includes Deep Learning, Computer Vision, Soft Computing, Cyber Physical Systems and Internet of Things. She is a member of Computer society of India and Soft Computing Research Society

Retrieval Number: 100.1/ijrte.A58540510121
DOI: 10.35940/ijrte.A5854.0510121

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication