IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

Wine Quality Prediction using ML Techniques and KNIME

Prasanna M[1] and Kamalesh Kumar S[2]
Students, School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology, Chennai, Tamil Nadu, India
prasanna.m.official@gmail.com[1] and skamaleshkumar25082000@gmail.com[2]

Abstract: *The Wine quality is important for purchasers as well as the wine industry to produce in good quantity. The normal way of quantifying wine quality is tedious. These days, machine learning models are key tools in replacing human tasks from measuring alcohol quality. While in quality prediction, there are several features, but not all the traits will not be relevant to quality prediction. Classification of wine quality is a complex work as the Flavour is the least aspect of human senses. For wine quality prediction RFC, SVM, Logistic Regression, GDC and Bayesian classifier demonstrates to be better with greater prediction accuracy than other data mining techniques. This prediction can be used in CART, SVM, Random Forest (RF) and Big-Data. The performance of the proposed model achieved the highest classification accuracy (99%) using Random Forest classifier. The paper explores which of the features wine determines the best quality of wine and generate insights into each of these features.*

Keywords: Quality of wine, Gradient descent classifier (GDC), Logistic Regression, Hadoop, Random Forest Classifier (RFC), Decision Tree Classifier (DT), Knime.

I. INTRODUCTION

Machine learning is a subset of AI that focuses on creating data-driven systems that improve their accuracy over time without being instructed. Prediction is the most significant data mining technique [1], which uses a group of pre-classified cases to create a model that can identify and classify the relationship between dependent and independent variables. Wine quality prediction is used to predict the quality of alcohol present in that for a given quality from the given dataset. The attributes used for predictions are fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulphur_dioxide, total_sulphur_dioxide, density, pH, sulphate. This technique is used eventually and helps in improving wine quality in industries. ML technique uses a training dataset to generate a model. A model is essentially a technique that uses individual weights and training variables to obtain a target value. As we have good and bad quality by (zero and one) to each variable for each record indicates the model how that variable is expounded to the target value which is wine quality. There should be enough quality of training information to work out the simplest attainable weights of all the variables. A model will find the test of the wine the appropriate result or quality of wine by learning with appropriate weights as precisely as possible. Huge data can be organized and analyzed for predicting the quality of wine in different datasets and the process of organizing and processing.

II. LITERATURE SURVEY

Mohit Gupta, [3], proves the usage of different ML models in predicting the quality of wine and the results are attached through different significant measures. Contribution of different independent variables showing up the final result is accurately illustrated. Yogesh Gupta [1], explores the usage of ML techniques such as Regression analysis, NN and SVM for product quality in two days. Here Regression analysis is used to determine the need of target variable on independent variables. Devika Pawar [2], An Robotic forecasting system is combined into a decision support system, to help in improving the speed and quality of the performance. In addition, the process of selecting feature might help to probe the brunt of the analytical tests. If that several achieved input variables are immensely applicable in predicting the quality of wine, although in formulating process some variables can be restrained, those data's can be used to improve the quality

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

of wine. Paulo Cortez1, Juliana Teixeira1, Antonio Cerdeira2. [11]. Wine quality is modelled with white vinho verde samples collected from the Minho region of Portugal covered by a regression approach that retains the order of the grades.

III. Related Work

3.1 Linear Regression

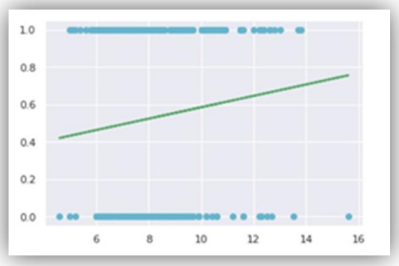
Linear Regression is a supervised learning technique that analyses the fixed acidity for predict wine quality prediction based on independent variables. This identifies the relationship between prediction and the dataset variables. This type of association that is evaluated between dependent and independent variables, as well as the number of independent variables, varies amongst regression models. The linear regression equation is computed using equation (1) where S and t are computed using equation 2 and 3.

$$\beta = s + t\alpha \quad [1]$$

$$s = \frac{(\sum \beta)(\sum \alpha^2) - (\sum \alpha)(\sum \alpha\beta)}{n(\sum \alpha^2) - (\sum \alpha)^2} \quad [2]$$

$$t = \frac{n(\sum \alpha\beta) - (\sum \alpha)(\sum \beta)}{n(\sum \alpha^2) - (\sum \alpha)^2} \quad [3]$$

where, α and β refers the regression line variables, $s \rightarrow y$-intercept, t → Slope, α → first dataset value, β → second data set.



3.2 Logistic Regression

Logistic regression technique is used when the inference are in categorical form. This method is being used for predicting the alcohol and residual present in that wine which defines the taste of the wine, Whether the wine is sweet, sour, salty or bitter. The Logistic regression is derived using equation 4

$$s = e\hat{}(x_0 + x_1 * t)/\left(1 + e\hat{}(x_0 + x_1 * t)\right) \quad [4]$$

where, s →predicted output, → bias or intercept, →coefficient for the single input value (t).

3.3 Naïve Bayes Classifier

Bayesian classifier is ML model which is credible and used for the task like classification. This is the one of the bayesian classification algorithm which is totally based on the concept of Bayes theorem i.e., the probability of A is happening when B has occurred could be found by multiplying the probability of B is happening when A has occurred and the probability of A. This is then divided by the probability of B.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad [5]$$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

Where $P(A|B) \rightarrow$ Posterior Probability.
$P(B|A) \rightarrow$ Likelihood
$P(A) \rightarrow$ Class Prior Probability
$P(B) \rightarrow$ Predictor Prior Probability

3.4 Decision Tree Classifier

Using Knime, DT presume instances by categorizing them down the tree from the root node to some leaf node, which contribute to the distribution of the instance. Any instance can be presumed by initiating at the root node of the tree, analyzing the target variable characteristic specified by root node, then moving down towards the tree branch node corresponding to the value of the attribute. The Subtree rooted at the new node is repeatedly processed by the classifier.
Entropy for 1 attribute:

$$E(S) = \sum_{i=1}^{c} -_{p_i} \log_2 p_i \quad [6]$$

Entropy for more than one attribute:

$$E(T, X) = \sum_{c \in X} P(c)\, E(c) \quad [7]$$

Information Gain(I):

$$I(T,X) = \text{Entropy}(T) - \text{Entropy}(T,X) \quad [8]$$

IV. PROPOSED WORK

The overall flow structure of the proposed Wine quality classification scheme is depicted in figure 1. The wine quality dataset taken from Kaggle [25]. Subsequently, the machine learning models where used which includes regression analysis and classification techniques to presume the better accuracy of the model. To classify given input data into good quality and bad quality, all data were fed into three different classifiers:

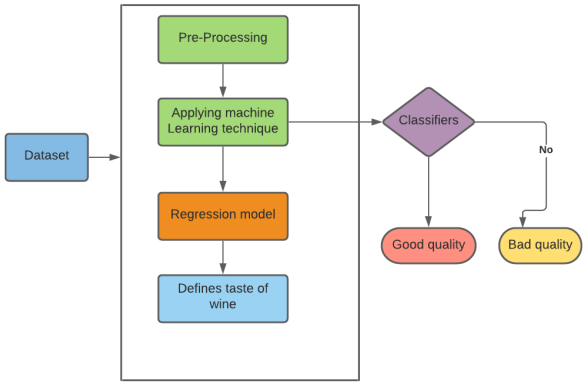


Figure 1: Proposed Wine Quality Classification model

The Hadoop distributed file system (HDFS) distributes, reserves, and facilitates access to massive amounts of data. The files reserved here are done in a redundant manner so that the system can be reused in the event of information loss, hence preventing failure. The HDFS allows parallel processing because the information is dispersed across several machines. The master server is the system with the name node, and it performs the following functions.

1. Controls the file system namespace.
2. Regulates client's ingress to files.
3. It also compiles the file system operations such as renaming, closing, and opening files and directories.

The data node is a hardware model with the Linux OS, where data node is installed. Each node includes a data node can be either hardware or system which are responsible for in data storage.

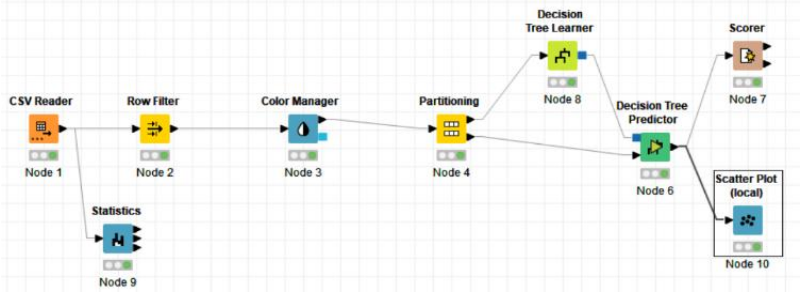
1. Data nodes respond to client requests by performing read and write operations inside the file systems.
2. They also achieve activities such as block create, delete, and replicate in according with some name node's instructions.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

In Knime, environment reads data from an excel file via "file reader" node by copying and pasting into the workbench, file is saved using .csv extension but before loading the data it has to be reevaluated to avoid: The central proclivity of info from that particular variable, with expected deviations. Follow the cleaning process, the info has to be uncovered to psychological measurement, if scales are utilized in the investigation.

1. The basic processing point of any data manipulations is a node.
2. The Order of steps or actions that is taken by the platform to achieve a particular task is a Workflow.

In decision tree classification, the node instigates a classification in main memory. Here we use either nominal or numerical attributes. It is that the Numeric splits we use are always binary, the domain is divided into two partitions at an apt split point. In other hand, Nominal splits can have as many outcomes as nominal values or can either be binary. In the case of a binary split nominal values are splitted into two subsets. The model provides two quality metrics for split calculation; the value of gain ratio and the Gini index. Further, there exist a post trimming method to reduce the tree size and increase prediction accuracy.



In this paper the quality of wine can be predicted using MapReduce programming and presumed that the accuracy of the quality falls between 95 to 99 % is achieved.

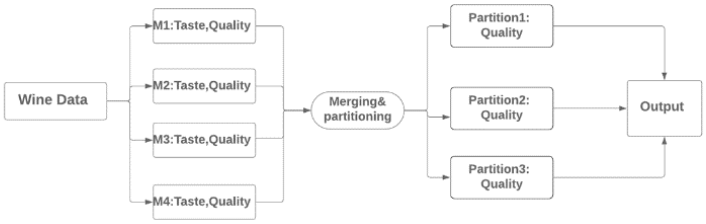


Figure 2: MapReduce Architecture

In this paper partitioned the quality of wine with taste from the wine data using MapReduce programming and the quality of wine falls partitioned into three parts such as bad, good and better is achieved.

V. Experimental Results and Discussion

Kaggle [25] included wine data where there are more than nineteen thousand observations with 14 attributes(columns). The attributes include are fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulphur_dioxide, total_sulphur_dioxide, density, pH, sulphate etc., After collecting the data the raw data is split into training and testing set where the target variable is the quality. The Training dataset contains the best quality of wine. A 10-fold cross-validation technique is executed to achieve consistent evaluation for classification accuracy on each classifier. The best performance to the classifier is assessed based on the average classification accuracy of the 10 folds. In this research work, Machine Learning models were used which includes Regression Analysis/Analytics which gives the accuracy of 89 % and Logistic regression which gives an accuracy of 99%. In the further analysis the data is classified as Good or bad quality which is trained with 80% and tested for 20%. The model used for classifying wine data are Support Vector Machine (SVM) which provides the accuracy of 69%, Gradient classifier which provides an accuracy of 98% and Random Forest Classifier (RFC) which gives a highest accuracy of 99.8%. The proposed algorithm's accuracy is compared to that of existing and is given in Table 1. Experimental analysis of RFC method proved to provide better

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

generalization ability in which defines the taste and quality of wine yielded higher classification accuracy when trained with all the classifiers. The Random Forest classifier model was set for training with 250 trees with each tree being built considering 4 random features.

Inference 1: Comparison of the proposed accuracy with the existing algorithm

Ref No	Prediction	Class	Attributes	Accuracy	Algorithm
[2]	Wine Quality Prediction using Machine Learning Algorithms	Binary	Residual sugar, fixed acidity and alcohol	87.33%	Random Forest Classifier
[30]	Predicting quality of wine based on chemical attributes	Binary	Alcohol, pH, density and quality	k=9 minimizes the cross validation residual mean squared error (RMSE).	K-Nearest Neighbour
[3]	A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality	Binary	Alcohol, pH, density and quality	81%	M5P
Proposed	Red White wine quality	Binary	Temperature, wind, and humidity	99%	Random Forest and Logistic Regression

The regression model is shown in figure 3(a) and (b) for Good and Bad Quality

(a) (b)

Fig 3(a) and (b): Linear Regression Analysis

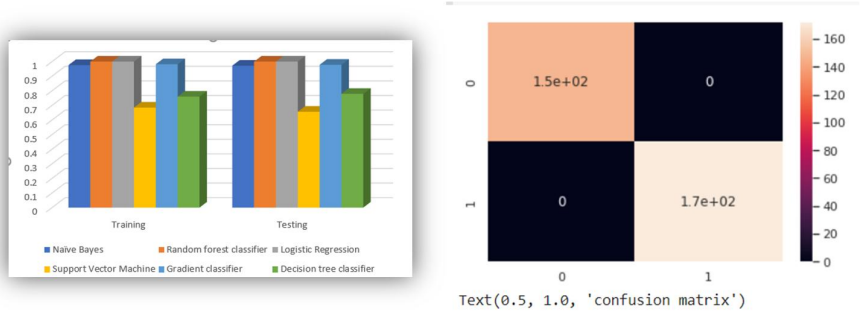
Similarly, the fixed acidity vs wine quality along with the average of alcohol and its residual values we have define the quality and taste for that wine present in our dataset.

(a) (b)

Figure 4(a) Quality by taste 4(b) PH Range by pH values

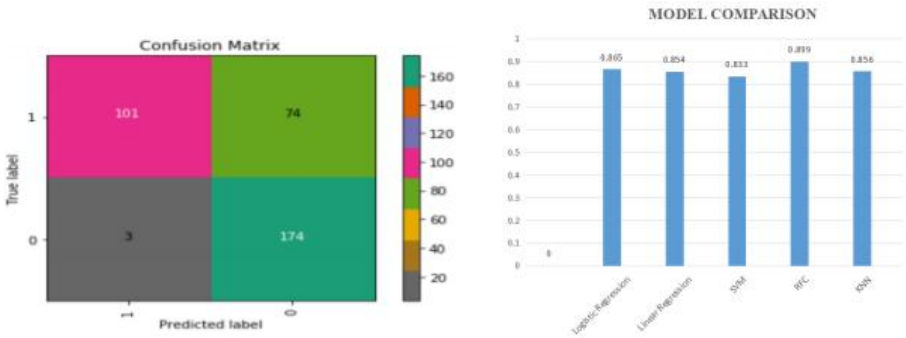
The performance of the Random Forest classifier is evaluated and achieved high classification accuracy when compared with the other classifiers. Based on the figure 4(a) good quality is more and bad is very less. In 4(b) PH Range of the strong acid contains the color of red compare to week acid. Experimental results confirm the potential of the proposed system to be used by the wine industry to predict the quality of wine.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252



(a) (b)

Fig 5: (a) Comparing Machine Learning Models 5(b)Confusion Matrix



(a) (b)

Fig 6(a): Confusion Matrix **6(b)** Comparative Analysis of Machine Learning models

VI. Performance Analysis

This section deals with various performance metrics used to analyze the wine quality and taste.

6.1 Accuracy

The criterion for evaluating classification models is accuracy. The precision is the percentage of the classification model's predictions that are correct. Accuracy is formulated as formulae (6),

Accuracy=((TP+TN)/(TP+TN+FP+FN)) **[6]**

where, TP refers the True Positive, TN refers True Negative, FP refers False Positive and FN refers False Negative.

6.2 Precision

Precision is the ratio of predicted positive result to the total predicted positive result. Precision is formulated using equation (7)

Precision = TP / (TP + FP) **[7]**

where TP → true positive and FP → false positive

6.3 Recall

Recall is the measure of finding the correct true positive values. Recall is formulated as equation (8)

Recall = TP / (TP + FN) **[8]**

where, TP → True Positive and FN → False Negative.

6.4 F1-score

F1 Score or F measure shows the balance between Precision and Recall. F1 Score is formulated as equation (9)

F1Score=2*((precision*recall)/(precision +recall) **[9]**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

VII. CONCLUSION

The Best wine quality is identified by its components like smell, taste and color. We can also predict the quality of wine using Machine Learning technique, we can help industries to certify the process of classifying whether it is good quality of wine or bad quality of wine. We have designed a model to predict the wine quality conditions using various ML techniques and Bigdata framework. This various algorithm gives us the different and best accuracy from the given dataset. The obtained outcome of the algorithm will be differentiated by different evaluation metrics like precision, recall, accuracy and confusion matrix. From the above observations, Finally, we presume that Random Forest Classifier and Logistic Regression gives the highest and best accuracy of 99% to predict the quality of wine. In future 100% accuracy is obtained by applying the big data and data Analytics techniques.

ACKNOWLEDGMENT

In the accomplishment of completion of our project on "**Wine quality prediction using ML techniques and knime**". We would like to convey our special gratitude our prof. Pattabiraman, of School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. Your valuable guidance and suggestions helped us in various phases of the completion of this project. we will always be thankful to you in this regard. We are ensuring that this project was finished by us and not copied from any other resources.

REFERENCE

[1]. Yogesh Gupta, "Selection of important features and predicting wine quality using ML models", 6th International Conference on Smart Computing and Communications, December 2017, Kurukshetra, India.

[2]. Devika Pawar, Aakanksha Mahajan, Sachin Bhoithe, "Wine Quality Prediction using ML Algorithms", International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 2019.

[3]. Mohit Gupta, Vanmathi C, "A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality", International Journal of Recent Technology and Engineering, Volume-10 Issue-1, May 2021.

[4]. Satyabrata Aich, Ahmed Abdulhakim Al-Absi, Kueh Lee Hui, Mangal Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques", 2019 21st International Conference on Advanced Communication Technology.

[5]. Gupta, Y., 2018. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science.

[6]. Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining from Physicochemical Properties. Decision Support Systems, Elsevier, 47,

[7]. 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".

[8]. Yunhui Zeng1, Yingxia Liu1, Lubin Wu1, Hanjiang Dong1. "Evaluation and Analysis Model of Wine Quality Based on Mathematical Model, Jinan University, Zhuhai,China.

[9]. Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) "Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception". Analytica Chimica Acta 484 (1).

[10]. Ebeler S. (1999) "Flavor Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine". Kluwer Academic Publishers, 409–422.

[11]. Paulo Cortez1, Juliana Teixeira1, Ant´onio Cerdeira2. "Using Data Mining for Wine Quality Assessment".

[12]. Yesim Er*1, Ayten Atasoy1. "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities", 3rd September 2011.

[13]. Shuhao Zhang, Caixing Shao, Wei Xiao, "Research on Red Wine Quality Based on Data Visualization", 2020 3rd International Conference on Artificial Intelligence and Big Data.

[14]. Larkin, T. and McManus, D. (2020) An Analytical Toast to Wine: Using Stacked Generalization to Predict Wine Preference. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, February 2022
Impact Factor: 6.252

[15]. P. Cortez, A. Cerderia, F. Almeida, T. Matos and J. Reis, "Modelling wine preferences by data mining from physicochemical properties", In Decision Support Systems Elsevier, vol. 47.

[16]. S. Ebeler, "Linking Flavour Chemistry to Sensory Analysis of Wine" in Flavor Chemistry Thirty Years of Progress, Kluwer Academic Publishers.

[17]. V. Preedy and M. L. R. Mendez, "Wine Applications with Electronic Noses" in Electronic Noses and Tongues in Food Science, Cambridge, MA, USA:Academic Press.

[18]. S. Kallithraka, IS. Arvanitoyannis, P. Kefalas, A. El-Zajouli, E. Soufleros and E. Psarra, "Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin", Food Chemistry, vol. 73.

[19]. N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer", Instrum. Measurement IEEE Trans, vol. 57.

[20]. S. Shanmuganathan, P. Sallis and A. Narayanan, "Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality", IEEE International Conference on Computational Intelligence Communication Systems and Networks, July 2010.

[21]. B. Chen, C. Rhodes, A. Crawford and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel", IEEE International Conference on Data Mining Workshop, Dec. 2014.

[22]. J. Han, M. Kamber and J. Pei, "Classification: Advanced Methods" in Data Mining Concepts and Techniques, Waltham, MA, USA:Morgan Kaufmann, 2012.

[23]. W. L. Martinez and A. R. Martinez, "Supervised Learning" in Computational Statistics Handbook with MATLAB, Boca Raton, FL, USA:Chapman & Hall/CRC, 2007.

[24]. K.Thakkar, J. Shah, R.Prabhakar, A. Narayan, A. Joshi, “AHP and machine Learning techniques for wine recommendataions”, International Journal of computer science and Information technologies,7(5).

[25]. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009