IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING APPROACH FOR ATTRIBUTE IDENTIFICATION AND QUALITY PREDICTION OF RED WINE

[1]Kushalatha.M.R., [2]Rachana Pai, [3]Sameep Pai. K.S. [4]Shrreesha.S. Bhat

[1]Assistsnt Professor, [2,3,4] Students
[1,2,3,4] Department of Electronics and Communication,
[1,2,3,4] Nitte Meenakshi Institute of Technology, Bangalore, India.

Abstract: The paper focuses on quality prediction analysis of red wine which contains the attributes such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfuric acids, total sulfuric acid, density, pH, sulphates, alcohol - quantity. The Prediction is evaluated through various machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree Classifier and Random Forest Classifier. By applying these algorithms, a detailed view of the tested results and the trained results of the machines were obtained. The results were visualized using Boxplot, Correlation matrix, joint plot. The Support Vector Machine (SVM) algorithm yields highest Accuracy of 87.5%, when used with grid search CV the accuracy was approximately equal to 90%.

***Index Terms* - SVM, SVM with GridSearch CV, Machine learning, Random Forest Classifier, Hyper parameters.**

I. INTRODUCTION

Nowadays the social drinking is on rise, which has shown exponential growth in the red wine industry. The red wine industries are making use of product quality certifications for the promotion of their products. The quality certification is bases on the comments given by experts which makes the quality certification process time consuming and also expensive. Due to the variations in the high degree of opinions among the wine tasters, there exists a variable degree of price difference. Along with these opinions, the laboratory based physiochemical tests also plays a vital role in determination of assessment of red wine quality. The factors used for the judgement of results are: acidity, pH level, sugar, and chemical properties. If the analysis of wine would have been based on human quality of tasting then, the analysis would have been controlled. The chemical properties that can be listed during the process of analysis gives good quality assessment. This paper lists all those best quality red wine indicators and features. The analysis also generates insights into each of these factors to the model developed for analysis of red wine quality.

II. DATASET

Various chemical information about the red wine can be extracted from the knowledge obtained from the data set. The red wine dataset procured has 1599 instances with 12 variables each. The main advantage of dataset is that they provide excellent results when used in classifiers for classification applications, also in regression analysis using regression models. The model with collected dataset is basically used in the analysis purpose of the quality of red wine. The analysis is incorporated by applying various algorithms like regression, decision tree, random forests. The basic advantage lies in differentiating the number of models used for the prediction and analysis purpose. The evaluation leads to performance differentiation and tabulation. Later stages of evaluation lead to generation of quality rating. The quality is evaluated based on the rank that can be assigned from sensory test. The quality rating starts with a scale of 0-10 class rates, wherein 0 is considered to be bad and 10 is considered to be excellent. The evaluation has a total of 11 classes assigned for evaluation purpose.

III. LITERATURE SURVEY

In the work carried out by Ribeiro et.al wine quality analysis prediction was done using data mining tools. The method was considered to be one of the best tools for measurement. The organoleptic parameters, which are obtained generally form the chemical components of the red wine was taken into consideration. Classifiers like Decision trees, Artificial Neural Network and

Linear Regression were made use for the evaluation purpose. The outcome of the paper showed good accuracies in all the techniques used [1]. In the research carried out by, Lee et.al proposed a decision tree-based method and was incorporated to predict the wine quality. The comparison of the obtained result was done with the existing methodology named as WEKA (based data mining tool). The WEKA approach was the hybrid approach where higher accuracy was possible by summing up as Support Vector Machine, Bayes Net and Multi Perceptron method. The Hybrid method resulted in better results in comparison with other approaches [2] with respect to the accuracy of classification parameters. For wine quality analysis Appalasamy et al., applied two classification algorithms, decision tree and Naive Bayes [3]. In a study by Gongzhu Hu, Tan X used Naïve bayes and decision tree by pre-processing the imbalanced data using SMOTE [4]. In a study made by Sruthi Pet.al, different samples of wine samples were collected. Various attributes that help to determine the quality assurance and different data mining classification algorithms such as Naive Bayes, Simple Logistic, KStar, JRip, J48 are applied on the data taken. The outcome of this work resulted in the classification of red wine into 3 main categories and a better accuracy compared to the previous study was obtained.[5] The work carried out by, V. Preedy and M. L. R. Mendez it was concluded that, as to get a good quality wine by providing less cost, it is must that the contribution of various chemical compositions and their concentration in a wine sample is essential. [6]. The wine data set used by these researchers are publicly available in the database of UCI. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. This data set collected always is a combination of variables related to components as well as related sensory attributes. The dataset collected by the researchers mainly focused on 12 attributes [7].

IV. RESEARCH METHODOLOGY

In the proposed work, machine learning techniques are used to obtain how wine quality is dependent on variables taken and also gives an insight into wine quality predictions. In stage 1 of our analysis, pre-processing of datasets are performed. Further, data model called as visualization model is obtained, which shows the amount of dependency of the wine quality on 11 independent variables that are considered for evaluation purpose(predictors). Then after, among the considered predictors foe evaluation, those predictors are taken into account which one feels that they are must to be taken. Based on the calculation which is based on to what extent the independent variables are dependent and also how the wine quality differs on these dependencies the analysis is carried out. At last, Wine quality is predicted with the help of support vector machine with GridSearch CV considering all predictors and selected predictors. The machine learning approach carried out uses the following approaches.

- **Logistic Regression:** Method of Logistic Regression works on the basis of assigning various observations to a discrete set of classes. The function called as logistic sigmoid function is the output of logistic regression, wherein in linear regression output is always a continuous function. The sigmoid function used in the designated method will return a probability value. The probability values are used for mapping of classes in future analysis purpose. The values will be for mapping of 2 or more classes.

- **The random forest (RF) classifier**: A supervised classification technique, is based on the concept of tree-based classifiers. Each tree is a pre trained with the data named as training data. The input to the classifier is the data or images that is to be classified and the measure to be taken to feed this input data/images to each tree present in the forest. Each tree is involved in classification process and is known by the term “vote” for that class. The forest having the highest vote is selected by the forest for the classification process. The determination of split is important in the classification process which is accompanied by a process of searching across a random subset of variables at each node in the tree. The Random forest classifier is extensively used due to its prominent characteristics like high accuracy and high speed of processing. In RFC care should be taken with respect to choosing the independence and correlation that actually affects the accuracy of the final cover map.

- **K- Nearest Neighbour Algorithm (KNN):** The proposed method consists of the following steps. Initial step is to adapt a support vector machine (SVM) approach. The SVM helps to obtain classification probability maps which in turn help to reflect the probability with which each of the pixels of the HSI belong to the different classes of classification. The filtering process of KNN includes matching and non-local averaging of the neighbourhood pixels. The method is proved to be a competitive method for classification as it doesn’t make use of and optimization strategies. The KNN when applied on two widely used data sets proved that classification efficiency is better compared to the traditional methods. The KNN method’s operating principle is based on a rule: majority voting rule, which presumes that all the neighbouring pixels are equally contributing to the classification process. Another important parameter used by KNN is Euclidian distance metric, which works as a distance metric when the data is homogeneous.

- **Spectral Angle Mapper (SAM):** SAM is one of the types of supervised classification technique used for classification of Hyperspectral data. The method makes use of the information regarding the spectral angle of an image, thus obtains the classification very quickly. The method is very quick as it determines the reference spectra from its field measurements. The spectral angle is the measure of a n- dimensional vector between the testing image and the reference spectrum with which the image is compared. The value of the angle determines the similarity and the difference between the image and the spectrum. If the angle is found to be larger, then the similarity is smaller and vice versa.

- **Naive Bayes:** Naïve Bayes is an optimization technique which is purely based on Bayes' rule. The assumption made in prior to the application of this algorithm is that the attributes considered will be completely independent even if the classes are already known for classification. Naïve Bayes calculates the posterior probability P (y | x) of each class y given an object x for the given object information. Naïve Bayes is a form of Bayesian Network Classifier based on Bayes' rule:
  P (y | x) = P(y) P (x | y)/P(x), with the predefined assumption.

- **Gradient Descent:** Gradient descent is an improved version of the optimization technique whose role is to optimize the function value by minimizing the cost function. Gradient descent is best suited whenever the situation comes wherein it becomes difficult to obtain the parameters with the help of analysis and most other basic optimization algorithms fail. In situations whenever one comes across a huge data, stochastic gradient descent is made use of. The updating of coefficients for each run is caried out and is performed for each training data set as to derive the optimization.

- **Support Vector Machine:** SVM is a very strong and prominent classification approach, which makes use of Kernel functions and works on the basis of theory of optimization. SVM is proved to a better technique of classification in environments where the image is represented in a greater number of spectral bands and availability of a smaller number of training samples. Figure 1 shows a SVM model in multidimensional space which represents different classes in a Hyperplane. The Hyperplane is generated in an iterative manner as to reduce the error in classification. The maximum Marginal Hyperplane (MMH) is obtained by SVM by dividing the given datasets into variety of classes.

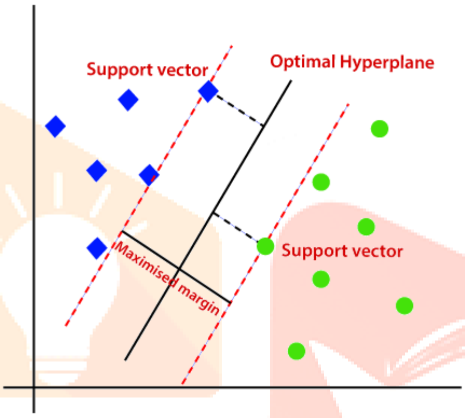

Fig 1: Hyperplane

- **Machine- Learning Model:** A Machine Learning model is a mathematical model which helps one to learn various parameters that is required to be learnt as to analyse the attribute calculation, required to analyse the quality of wine. However, Hyperparameters cannot be learned by the Machine learning model directly. These models show high performance in their analysis and evaluation part exhibit their importance by improving performance and is achieved by its complexity or its learning rate. SVM technique is also based on creation of Hyperplanes and decision function, Machine learning technique works with the principle of creation of a grid of hyper-parameters so that all the combinations can be tried out using the model.Scikit-learn has this functionality built-in with GridSearch CV. GridSearch CV takes a dictionary that describes the parameters that could be tried on a model to train it. The grid of parameters is defined as a dictionary, where the keys are the parameters and the values are the settings to be tested.

V. Results And Discussion

Accuracy, Mean Square Error [MSE], Root Mean Square Error [RMSE], ROC_AUC_SCORE [RAS] obtained for different machine learning model is given below:

Table 1: Result analysis

Model	Accuracy (in %)	MSE	RMSE	RAS
Logistic regression	86.56	0.134	0.366	0.6218
Decision tree classifier	87.50	0.125	0.3534	0.7506
Random forest classifier	89.5	0.0843	0.2904	0.7744
Stochastic gradient descent classifier	83.12	0.1687	0.4107	0.4959
Naïve bayes classifier	84.68	0.1531	0.3913	0.8221
K nearest neighbor classifier	87.81	0.1218	0.3491	0.6819
Support vector machine	87.50	0.1250	0.3535	0.6185
Support vector machine (Grid search CV)	89.68	0.1031	0.3211	0.6753

From the table 1, it can be inferred that SVM with GridSearchCV also called as SVM with hyperparameter tuning yields best accuracy. We conclude that SVM with hyperparameter tuning as the best model even though random forest is yielding less error because the hyperparameters are also taken into consideration. The ROC-AUC curve is as shown in figure 2. The confusion matrix used here is to compare the actual target values obtained with the one which is predicted through the mechanisms incorporated using the machine learning model as shown in figure 3.

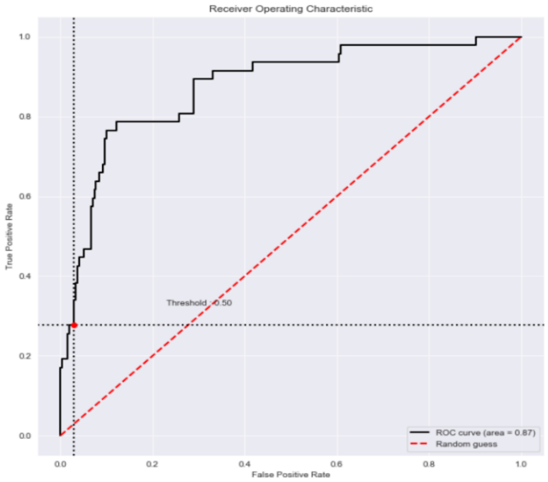

Fig.2: ROC-AUC Curve

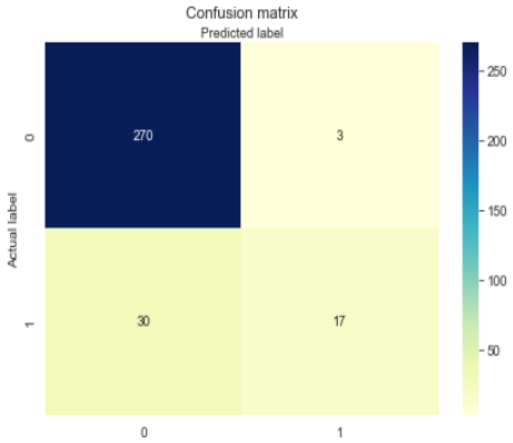

Fig.3: Confusion Matrix

The figure 2 shows the ROC AUC score for the SVM-Grid search CV model and the threshold was set at 0.5. AUC - ROC curve is a seen as a performance measurement evaluation parameter used in the classification problems. These evaluations are assigned with various thresholds. ROC is a known as the curve of probability and AUC indicates the degree or measure of separability between the various classes. The parameter shows the degree of capability of the proposed model in distinguishing the classes. The ROC curve is plotted with TPR against the FPR where TPR is plotted on the y-axis and FPR is plotted on the x-axis. From the results, the obtained value of AUC is 0.87, which in turn indicates that that 87% chance that the model will be able to distinguish between positive class and negative class.

Totally 12 variables in the dataset considered. The dependent variable is the quality of rating and remaining 11 variables out of considered 12 variables are considered as predictors or independent variables in the proposed model.

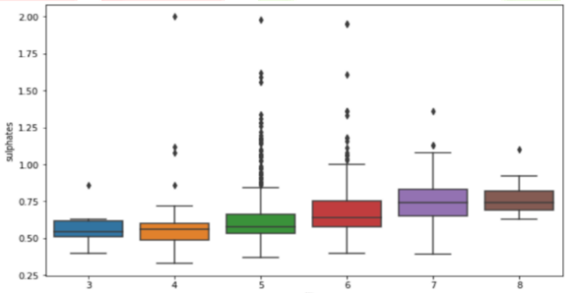

Fig.4: Quality Vs Sulphates

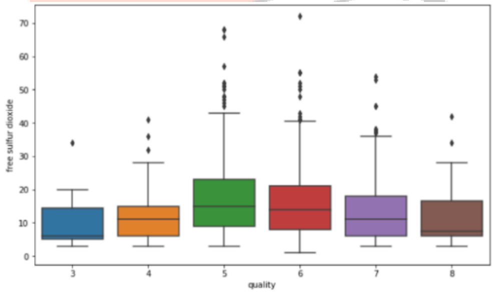

Fig.5: Quality Vs Free Sulfur Dioxide

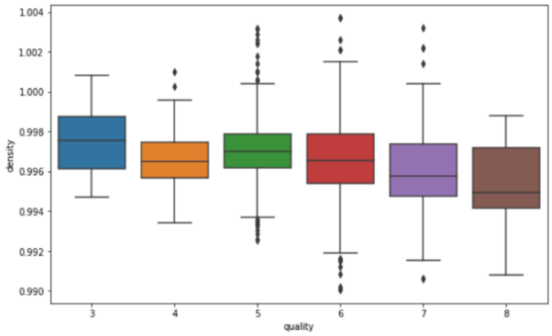

Fig.6: Quality Vs Density

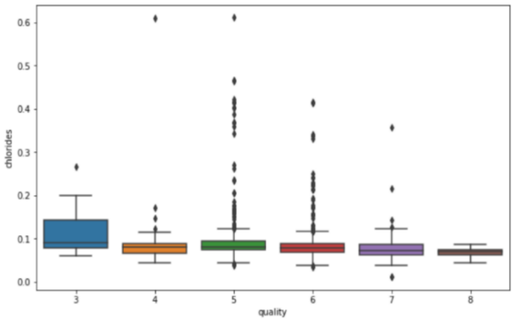

Fig.7: Quality Vs Chlorides

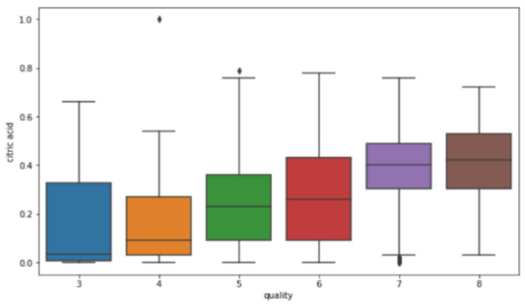

Fig.8: Quality Vs Citric Acid

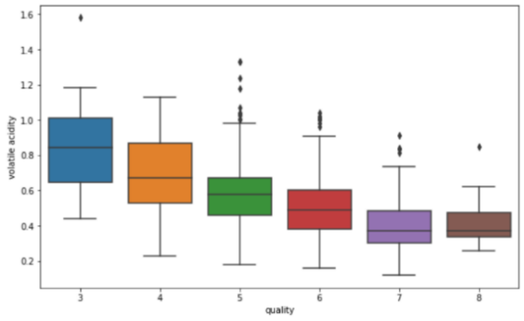

Fig.9: Quality Vs Volatile Acidity

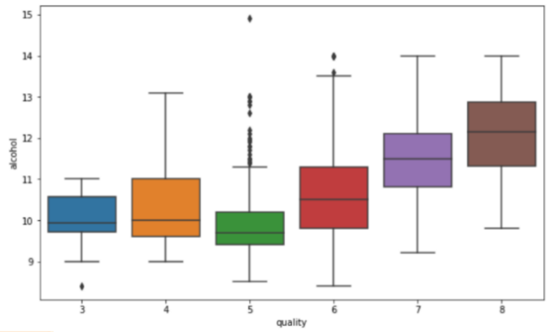

Fig.10: Quality Vs Alcohol

Figure 4 to Figure 10 depicts the plot of Quality versus features that contributes over the quality. The features which led to better prediction of good wine quality were volatile acidity, citric acid content, chlorides, sulfur-di-oxide, density, sulphate levels, alcohol.

VI. CONCLUSION

The paper focuses mainly on two types of analysis: firstly, the level of priority that is been assigned to each predictor which helps in evaluation of the wine quality is identified and secondly, the value assigned to wine quality which is obtained with the help of features extracted. SVM with GridSearch CV has been chosen as one of the best models among the models implemented. The Support Vector Machine (SVM) algorithm yields highest Accuracy of 87.5% which also gives a good accuracy of 89.6875% when used with GridSearch CV. As a further work better estimators or ensemble methods can be used with random forest and other models and then test for better results.

REFERENCES

[1] Ribeiro, Neves, Sanchez, Delgado, Machado and Novais, "Wine vinification prediction using data mining tools," 2009.

[2] Lee, Park, Kang, "Assessing wine quality using a decision tree," 2015.

[3] P Appalasamy, A Mustapha, N. D. Rizal, F Johari, and A. F. Mansor, "Classification-based data mining approach for quality control in wine production," 2012.

[4] Gongzhu Hu, Tan Xi, Faraz Mohammed, "Classification of Wine Quality with Imbalanced Data," 2006.

[5] Shruthi P, "Wine Quality Prediction Using Data Mining”," 2019.

[6] V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," 2017.

[7] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, "Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems," 2009.

[8] I. Janszky, M. Ericson, M. Blom, A. Georgiades, J. O. Magnusson, H. Alinagizadeh, and S. Ahnve, "Wine drinking is associated with increased heart rate variability in women with coronary heart disease," 2005.

[9] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020.

[10] S. Aich, A. A. Al-Absi, K. Lee Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," 2019.

[11] G. Hu, T. Xi, F. Mohammed and H. Miao, "Classification of wine quality with imbalanced data," 2016.

[12] S. Zhang, C. Shao and W. Xiao, "Research on Red Wine Quality Based on Data Visualization," 2020.