Prediction of Wine Quality and Classification

Valicherla, Sai Prudhvi
Department of Computer Science
svalicherla1@student.gsu.edu
Georgia State University
Atlanta,GA 30302

***Abstract*—Many companies have promoted their products in recent years based on the product's quality certification. Since taste is the least understood by the human senses, wine classification is a challenging activity. A good wine quality forecast can be handy in the certification process because humans currently conduct sensory analysis, which is a subjective process. Traditional methods of assessing product quality take time, and data mining is the best method for accomplishing this because it analyzes the data set and collects valuable information. With the advent of machine learning techniques, the systems have become more efficient and take less time. In this article, I have focused on a few machine learning strategies. A decision support system may incorporate an automated predictive system to improve the performance and efficiency of the process.**

keywords - Wine Prediction, Data Mining, Machine Learning, Decision support system.

I. Introduction

There has been a slight rise in wine consumption in recent years, and it has been observed that wine consumption has a positive association with heart rate variability [7]. With the increase in wine demand, the wine industry is searching for new ways to manufacture high-quality wine at a lower cost. Wines vary in color, vintage, ingredients, taste, aging process, and alcohol content, among other things. Understanding wine's distinct characteristics and how these characteristics contribute to their quality (or grade), price, and origin has proven to be a difficult task. While much research has been conducted on wine's physical and chemical properties, extracting objective data from subjective and human-dependent sensory data has proven to be more difficult.

While most of the chemicals are the same for various types of wine based on chemical tests, the amount of each chemical has a different concentration level of each kind of wine. It is essential to distinguish other wines these days to ensure consistency [8]. Due to a lack of technical resources in the past, it was difficult for most companies to identify wines based on chemical analysis because it took a long time and cost a lot of money.

Data mining is the process of uncovering new examples to separate high-quality data from a massive database. It includes different types of performance metrics, machine learning, and database organization. The main goal is to extract significant knowledge from an enormous database and transform the critical substance into something useful for future analysis.

In data mining, we use factual models and machine learning. Data mining is a computer-assisted method of extracting patterns from large quantities of data, identifying inconsistencies, and eventually getting the desired result. Various data mining algorithms and their best features are stitched together for better performance, resulting in fewer errors and reliable results.

Machine learning techniques have made it feasible to classify wines and assess the importance of each chemical composition in the wine and which ones to ignore for improved quality.

II. Background and Related Work

Various machine learning methods and feature selection techniques have been applied to the wine dataset in the past. Chen, Rhodes et al. suggested a method for predicting wine grade based on savory human feedback. They used a hierarchical clustering method and an association rule algorithm by processing the reviews to predict the wine quality and obtained an accuracy of 85.25 percent [3]. Thakkar, Shah et al. ranked the attributes using the analytical hierarchy approach, using support vector machine and random forest machine learning classifiers. They obtained 70.33 percent for random forest and 66.54 percent for SVM, respectively [4]. Appalasamy, Mustapha et al. suggested a method for predicting wine quality based on physicochemical tests. They say that using a classification system helps to improve wine quality during production [5]. To suggest the product, Reddy and Govindarajulu used a user-centric clustering approach. For the survey, they used the red wine data collection. Based on the literature review, they assigned relative voting to the attributes. They then used the Gaussian Distribution Process to give weight to the features. They assessed the standard utilizing the consumer preference category as a criterion [6].

Because of the previous work, I have decided to compare the performance metrics of various feature selection algorithms and classifiers. This paper used different classifiers such as Gradient Booster, Simple tree(decision tree) algorithm, and random forest tree. It used an Extra tree classifier and chi-square score-based feature selection.

III. Methodology

A. Data Preparation

For Wine Quality Prediction, the data is extracted from the UCI machine learning repository [1]. The dataset contains 1599 red wine instances and 4898 white wine instances, each with 12 variables, including 11 physicochemical variables

that affect wine quality. These two datasets are related to red and white "Vinho Verde" wine from Portugal. For Wine classification, the data is extracted from the UCI machine learning repository [2]. The dataset consists of 178 instances with 14 variables with 13 physicochemical variables that determine the Wine type.

The data is evaluated based on the inputs given and predicts the wine quality and wine type at the end. Wine contains tartaric, citric, and malic acids. While ascorbic, sorbic, and sulfurous acids are commonly added during the winemaking process. Residual sugar determines the sweetness of a wine. Although it is not the only factor determining sweetness, it is still essential in deciding a wine's flavor. Alcohol is a by-product of yeast metabolism in wine.

The data set's quality variable ranges from 3 to 9, with 3 indicating the worst quality and 9 indicating the best. Values 1, 2, and 10 are unusually absent. The quality values greater than or equal to 7 are considered High Quality, whereas the rest considered Low Quality.

Figure 1 shows the distribution of the instances in the data set by quality class for White Wine, while Figure 2 shows for Red Wine. Figure 3 shows the distribution of wine type values in the data set.

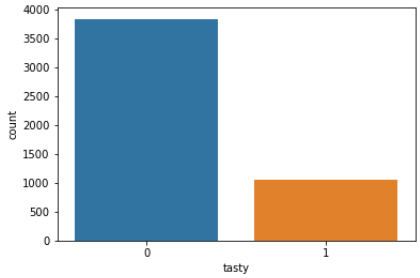

Fig. 1. Distribution of White Wine Quality values 0-low Quality 1- High Quality.

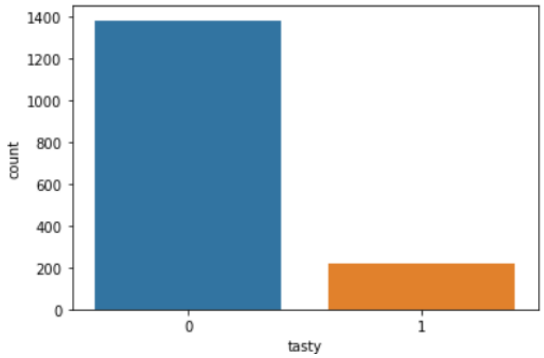

Fig. 2. Distribution of Red Wine Quality values 0-low Quality 1- High Quality.

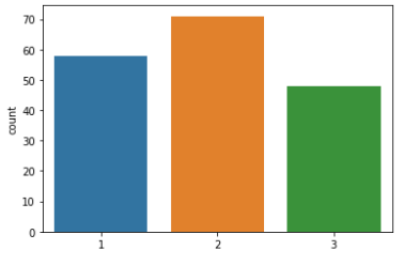

Fig. 3. Distribution of Wine types 1,2,3 in Wine data.

B. Flow Chart

Figure 4 shows the process flow from data preparation to predicting the results.

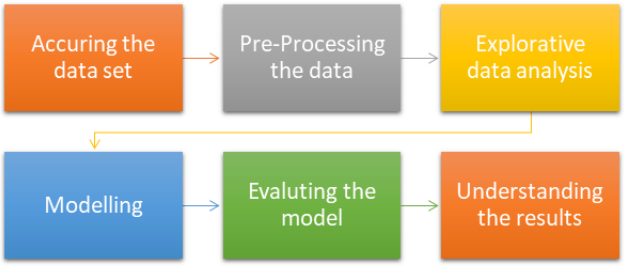

Fig. 4. Flow chart of Modelling

C. Feature Selection Methods

The process of filtering and screening the number of input variables to reduce the size and computation cost involved in developing a good prediction model is called Feature importance.There are different techniques involved to extract the best features.

1) Correlation:
The correlation matrix between variables gives the number of the features that are either dependent or independent on each other and the target variable.

TABLE I
WHITE WINE

Correlation	Feature
Strong positive	Density and Residual sugar (0.84)
Moderate positive	Density and total Sulphur dioxide (0.53), Total sulphur dioxide and free Sulphur dioxide (0.62)
Strong negative	Alcohol and density (-0.78)

TABLE II
RED WINE

Correlation	Feature
Moderate positive	Total sulphur dioxide and free sulphur dioxide (0.67), fixed acidity and citric acid (0.67), density and fixed acidity (0.67)
Strong negative	Alcohol and density (-0.5)

Based on the values in the correlation matrix, I have deduced a function to get the values of the features

which cross a certain threshold. After thorough research, and have finalized to set the threshold to be 90 percent for all data sets. From table 1,2,3 surprisingly I have found the none of the values have a higher correlation between each other for the given threshold.

TABLE III
WINE TYPE

Correlation	Feature
Strong positive	Flavonoids and od280od315 (0.79), Total Phenols and od280od315(0.70), Flavonoids and Total Phenols(0.86)
Moderate positive	DFlavonoids and proanthocyanins (0.67), Proline and Alcohol (0.64), Alcohol and colorIntensity (0.55)
Negative positive	Malic acid and Magnesium(-0.049), Ashalcalinity and Magnesium (-0.072), Alcohol and Hue (-0.075)

TABLE IV
APPLYING THE SELCTKBEST METHOD USING THE CHI2 SCORE TO RANK THE FEATURESAPPL

Features	CHI-SCORES
proline	16446.895876
colorIntensity	109.022694
flavanoids	62.991596
magnesium	41.953227
ashalcalinity	28.676542
malicAcid	27.896849
od280$_o d315$	23.022222
total phenols	15.531400
proanthocyanins	9.179720
alcohol	5.349870
hue	5.178420
non Flavanoid Phenols	1.794852
ash	0.742062

2) Select K best using Chi score:
In mathematics, the Chi-squared test is used to determine if two incidents are independent or not. We can use it in feature selection to see whether the frequency of a specific feature and the target are dependent or not.
3) Extra Tree:
It is a machine learning algorithm that incorporates the predictions of several decision trees into a single prediction. Bagging and random forest are also similar to the Extra Trees ensemble, a set of decision trees. The Extra Trees algorithm uses the testing data set to generate a massive number of unpruned decision trees. In the case of regression, predictions are made by averaging the projection of the decision trees, and in the case of classification, majority voting is used [14].

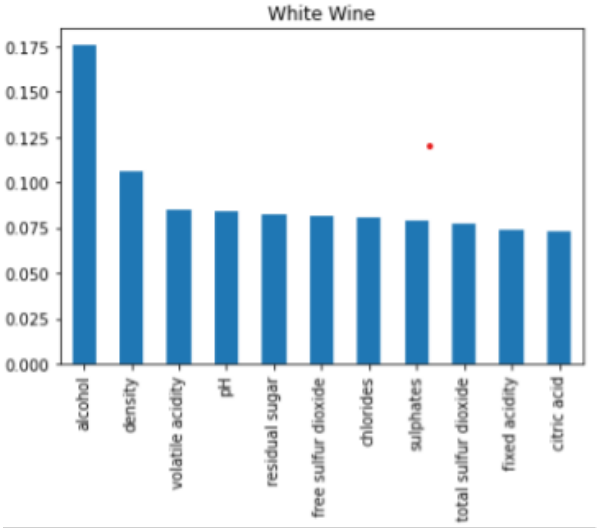

Fig. 5. Feature importance for the White Wine data using the Extra Tree classifier.

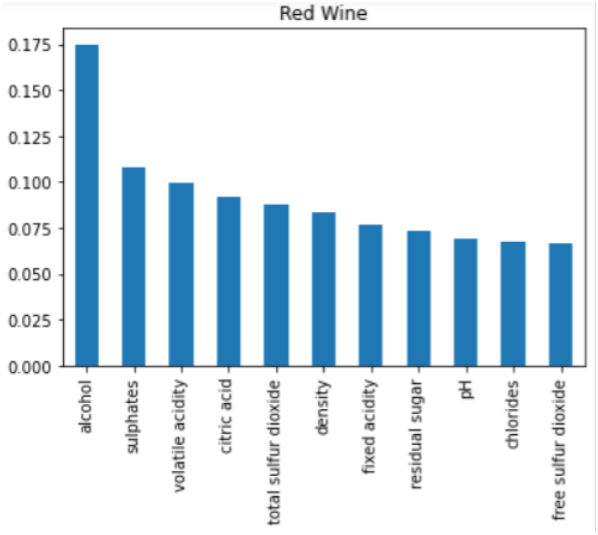

Fig. 6. Feature importance for the Red Wine data using the Extra Tree classifier.

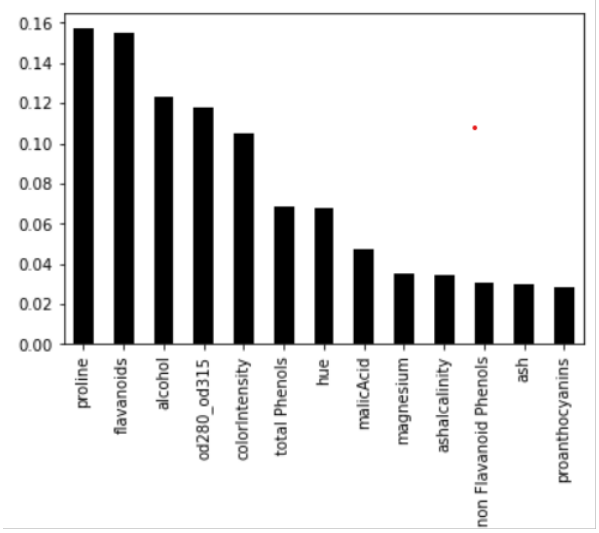

Fig. 7. Feature importance for the Wine classification using the Extra Tree classifier.

I have used the computed correlation function, L1,L2 regularization, Extra Trees Classifier, selectKbest using the chi2 score and found promising results while using the correlation function, selectKBest, and Extra Trees Classifier for the data set I have considered.

From the table 3 and 4, I have concluded that the features 'ash' does not contribute much to the target variable, and have decided to drop the feature.

D. Feature Scaling

Standard Scalar subtracts the mean from the variable and scales them to unit variance.

Normalization changes the values to a common scale without distorting the difference in the ranges in the values.

E. Classification Models

1) Gradiant boosting algorithm:
   Boosting is a technique for turning weak into good ones. Each new tree in boosting is based on an updated version of the original data set. The gradient boosting algorithm (gbm) is best demonstrated by first understanding the AdaBoost Algorithm. The AdaBoost algorithm starts by training a decision tree of equal weights for each observation. Following the evaluation of the first tree, raise the weights of the difficult-to-classify observations and decrease the weights of the easy-to-classify observations. As a result, the second tree is based on the weighted results. The goal here is to build on the first tree's predictions. As a result, our current model is Tree 1 + Tree 2.
   The classification error from this updated 2-tree ensemble model is then computed, and a third tree is grown to estimate the revised residuals. This method is repeated for a finite number of iterations. Following trees assist in the classification of findings that were not well classified by previous trees. The weighted total of the predictions made by the previous tree models makes up the final ensemble model's predictions. The way AdaBoost and Gradient Boosting Algorithm recognize the flaws of vulnerable learners is the main distinction between the two algorithms. Although the AdaBoost model uses high weight data points to identify flaws, gradient boosting uses gradients in the loss equation to do the same [9].
2) Decision Tree(Simple Tree) Algorithm:
   To determine whether to divide a node into two or more sub-nodes, decision trees employ various algorithms. The homogeneity of the resulting sub-nodes improves with the construction of sub-nodes. To put it another way, the purity of the node improves as the target variable grows. The decision tree divides the nodes into sub-nodes based on all possible variables and chooses the split that produces the most homogeneous sub-nodes [10].
   The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking. As the name suggests, a greedy algorithm always makes the choice that seems to be the best at that moment [11].
3) Random Forest algorithm:
   Ensemble learning is a type of machine learning in which many machine learning algorithms are combined to improve predictive accuracy. Random Forest is an example of ensemble learning [12]. For constructing trees, a random sampling of the training data set is used. When separating nodes, random subsets of features are considered. To produce an ensemble of trees, a technique known as bagging is used, in which several training sets are created with replacement. A data collection is separated into N samples using the bagging technique, which uses randomized sampling. The model is then based on all samples using a single learning algorithm. After that, the forecasts are mixed in tandem using voting or averaging [13].

IV. Experimental Results

I have used the following metrics to evaluate the performance of the model.

True Positive(TP): Predicted positive for an actual positive value.

True Negative(TN): Predicted negative for an actual negative value.

False Positive(FN): Predicted positive for an actual negative value.

False Negative(FN): Predicted negative for an actual positive value.

Precision: It is the ratio between the correct positives to the total positive predictions.

$$Precision = TP/(TP + FP)$$

Recall: It is the ratio between the correct positives and samples that should be positive.

$$Recall = TP/(TP + FN)$$

F-scores: Measures the model's accuracy on the data sets. It is the harmonic mean between precision and recall.

$$F1 - scores = (2 * precision * recall)/(precision + recall)$$

Accuracy: Accuracy is the machine learning model's performance metric and is defined as the percentage of correctness in the predicted value of the test data.

In other words, it is the ratio between the correct predictions to the total number of predictions.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

ROC stands for receiver operating characteristic curve. It graphs the performance of the model at all the classification

thresholds. The two parameters involved are True Positive Rate is the recall.

$$Recall = TruePositiveRate = TP/(TP + FN)$$

False Positive Rate FPR is a count of incorrect positive results among all negative samples during the test.

$$FalsePositiveRate(FPR) == FP/(FP + TN)$$

The area under the ROC curve (AUC): Measures the two-dimensional area under the ROC curve, and the value ranges from 0 to1. AUC is a metric that averages the overall classification thresholds performance. The likelihood that the model scores a random positive example higher than a random negative example is one way to view AUC. AUC is classification-threshold-invariant- It assesses the accuracy of the model's predictions regardless of the classification threshold used. AUC is scale-invariant- Rather than measuring absolute values, it assesses how well predictions are ranked.

A. White Wine Quality Prediction

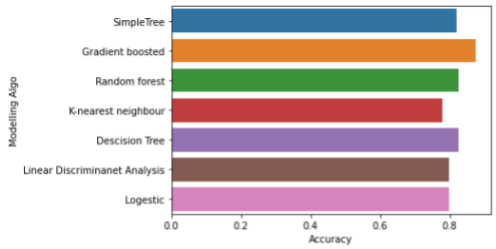

Fig. 8. Accuracies of different modelled Algorithms for White Wine data.

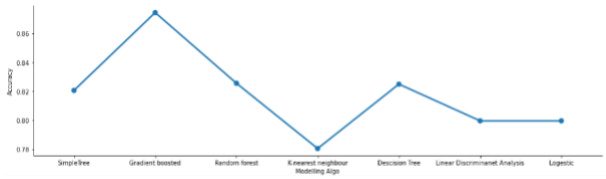

Fig. 9. Displaying Plot of all Accuracies for White Wine data.

I have found from the figure 9 that the Gradient Boosted algorithm has the maximum accuracy and the least accuracy is with the K-nearest neighbor.

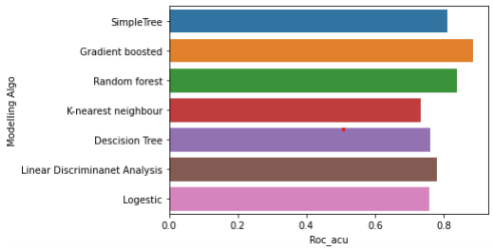

Fig. 10. Displaying ROC Curve for White Wine Data.

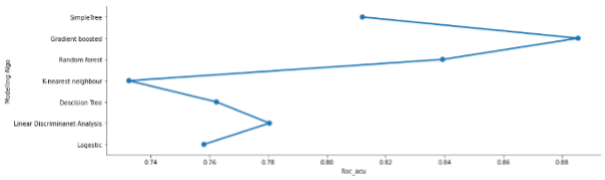

Fig. 11. ROC curve Plot for White Wine data.

I have found from the figure 11, that the Gradient Boosted classifying algorithms have the maximum AUC value, and the worst performance is with the K-nearest neighbor.

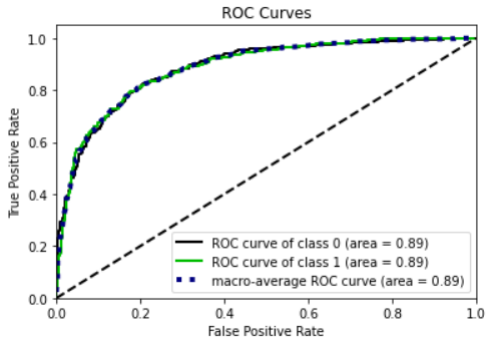

Fig. 12. ROC curve plot using Gradient Boost classifying algorithm.

B. Red Wine Quality Prediction

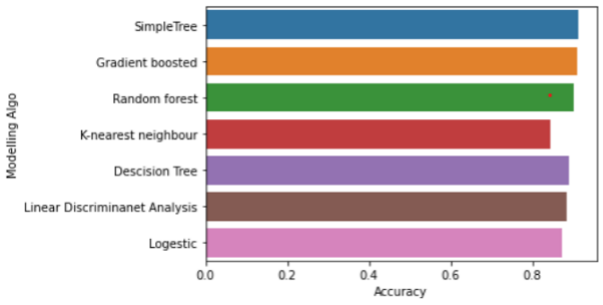

Fig. 13. Accuracies of different modelled Algorithms for Red Wine data.

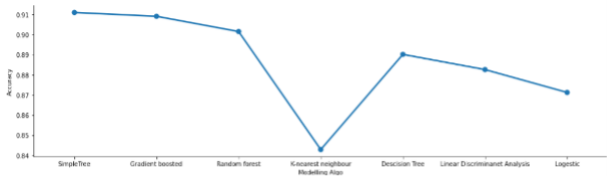

Fig. 14. Displaying Plot of all Accuracies for Red Wine data.

I have found from the figure 14, that the Simple Tree ad Gradient Boosted algorithm has the maximum accuracy and the least accuracy is with the K-nearest neighbor.

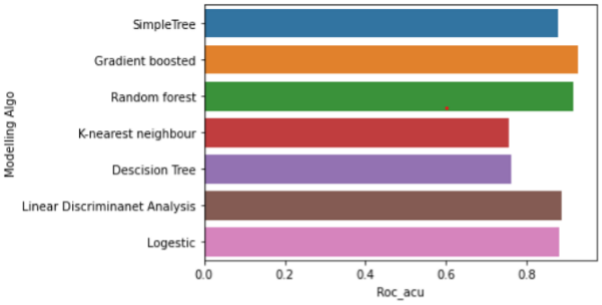

Fig. 15. Displaying ROC Curve for Red Wine Data.

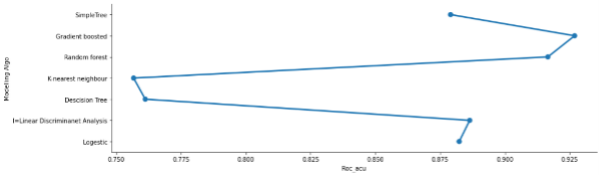

Fig. 16. ROC curve Plot for Red Wine data.

I have found from the figure 16, that the Gradient Boosted classifying algorithms have the maximum AUC value, and the worst performance is with the K-nearest neighbor.

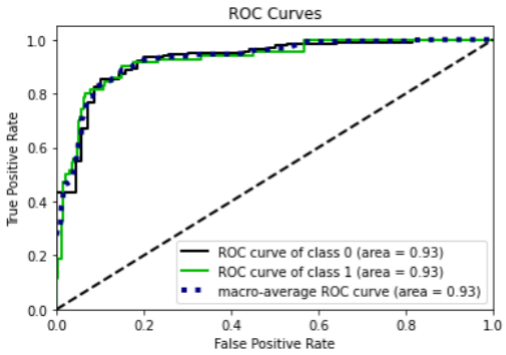

Fig. 17. ROC curve plot using Gradient Boost classifying algorithm.

C. Wine Classification

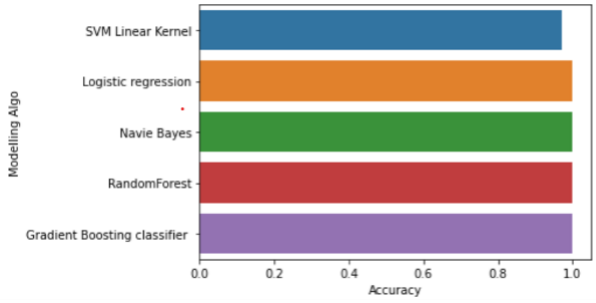

Fig. 18. Accuracies of different modelled algorithms.

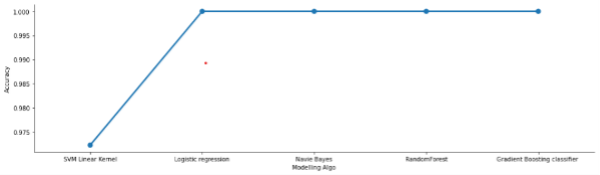

Fig. 19. Displaying Plot of all Accuracies for Wine Classification.

I have found from figure 19, that Logistic Regression, Navie Bayes, Random Forrest and Gradient Boosting classifier have the best performance and the least performance is observed in The Support Vector Machine with Linear Kernel suggesting that the data is not linear.

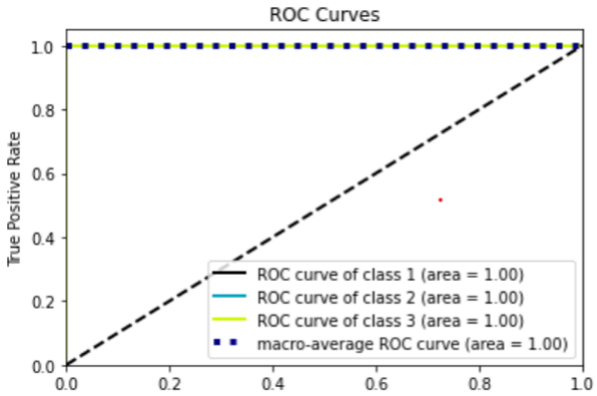

Fig. 20. ROC curve plot using Gradient Boost classifying algorithm.

V. Conclusion and Future Directions

Wine is considered a predominant horticultural product and is the best refreshment with many health benefits. So, the question that haunts us is, are we drinking the desired quality of wine? I have looked at the latest technology and used machine learning techniques to do the job for us. For classifying the wine class, the algorithms Logistic regression, Naïve Bayes, Random Forest, and Gradient Boosting produced a 100 percent accuracy, while the Support Vector Machine with Linear Kernel yielded an accuracy of 97 percent for the considered data set. Coming to predicting the quality of the white wine, Gradient Boosted Classifier generated an accuracy of 87 percent, and the least accuracy score is 78 percent for the K-Nearest Neighbour. On the other hand, for the Red wine data the Simple tree algorithm generated the highest accuracy of 91 percent, and K-Nearest Neighbour has the least performance accuracy of 84 percent.

In the future, our focus will be on predicting the cost of the wine using deep learning techniques based on the ingredients used and the time frame it was prepared. The longer the wine is left undisturbed, the tastier it will be and the costlier it is going to be.

References

[1] "UCI Machine Learning Repository", Wine quality data set, [online] Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.
[2] "UCI Machine Learning Repository", Wine data set, [online] Available: https://archive.ics.uci.edu/ml/datasets/wine.

[3] 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, pp. 142-149, Dec. 2014.

[4] K. Thakkar, J. Shah, R. Prabhakar, A. Narayan and A. Joshi, "AHP and MACHINE LEARNING TECHNIQUES for Wine Recommendations", International Journal of Computer Science and Information Technologies, vol. 7, no. 5, pp. 2349-2352, 2016.

[5] P. Appalasamy, A. Mustapha, N. D. Rizal, F. Johari and A. F. Mansor, "Classification-based Data Mining Approach for Quality Control in Wine Production", Journal of Applied Sciences, vol. 12, no. 6, pp. 598-601, 2012.

[6] Y. S. Reddy and P. Govindarajulu, "An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set", IJCSNS, vol. 17, no. 10, pp. 103, 2017.

[7] Janszky, M. Ericson, M. Blom, A. Georgiades, J. O. Magnusson, H. Alinagizadeh, et al., "Wine drinking is associated with increased heart rate variability in women with coronary heart disease", Heart, vol. 91, no. 3, pp. 314-318, 2005.

[8] 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, pp. 137-151, 2016.

[9] J. Son, I. Jung, K. Park and B. Han, "Tracking-by-Segmentation with Online Gradient Boosting Decision Tree," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3056-3064, doi: 10.1109/ICCV.2015.350.

[10] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - a survey," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 35, no. 4, pp. 476-487, Nov. 2005, doi: 10.1109/TSMCC.2004.843247.

[11] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.

[12] W. L. Martinez and A. R. Martinez, "Supervised Learning" in Computational Statistics Handbook with MATLAB, Boca Raton, FL, USA:Chapman Hall/CRC, pp. 363-431, 2007.

[13] J. Ham, Yangchi Chen, M. M. Crawford and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," in IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 492-501, March 2005, doi: 10.1109/TGRS.2004.842481.

[14] Manizheh Ghaemi, Mohammad-Reza Feizi-Derakhshi,"Feature selection using Forest Optimization Algorithm, Pattern Recognition",Volume 60,2016,Pages 121-129,ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2016.05.012.