Wine Quality Testing by using Machine Learning

| *Ramanjali Medarametla, 700747141*  *Computer Science*  *University of Central Missouri*  *Lee’s Summit*  *rxm71410@ucmo.edu* | Naga Chetan Kumar Reddy | Naga Chetan Kumar Reddy | Sridhar Seepana |
| --- | --- | --- | --- |
| *Mounika Nelluri, 700745442*  *Computer Science*  *University of Central Missouri*  *Lee’s Summit*  *mxn54420@ucmo.edu* | *Computer Science*  *University of Central Missouri* | *Computer Science*  *University of Central Missouri* | *Computer Science*  *University of Central Missouri* |
| *Sowmya Myla, 700741181*  *Computer Science*  *University of Central Missouri*  *Lee’s Summit*  *sxm11810@ucmo.edu* | Lee’s Summit  [cxn34080@ucmo.edu](mailto:cxn34080@ucmo.edu) | Lee’s Summit  [cxn34080@ucmo.edu](mailto:cxn34080@ucmo.edu) | Lee’s Summit  [sxs29730@ucmo.edu](mailto:sxs29730@ucmo.edu) |

***Abstract*—** **The most consumed alcohol is wine. Wine quality is being tested as part of our initiative. The quantity of years has an impact on wine quality. Wine quality is determined by a number of factors, including acidity, alcohol content, and fragrance components, among others. Each of these factors has an absolute value. We evaluate them using their standards, test them, and provide a report on the wine's quality. We are testing using machine learning We are utilizing the Random Forest Technique on supervised data.**

***Index Terms*—Machine Learning, Wine, Prediction*.***

1. INTRODUCTION

The wine business is advancing its technologies for both wine manufacture and sales in response to the growing consumer interest in wine and its adoption in new areas. Definitive

Although it is notoriously difficult to make claims about the quality of a wine, more unbiased measures are essential for growth into a larger market because new wine lovers want assurance on the value of a wine[1].

Wine quality is often evaluated using sensory and physiochemical techniques. Currently, a trained panelist's taste produces more accurate results, but this approach of gauging quality is time- and money-consuming[2]. In order to characterise wine, physical and chemical tests are frequently done. These tests examine the wine's density, sugar, tannin, alcohol content, and acidity levels.

1. MOTIVATION

There are several motivations for using machine learning in wine quality testing:

1. Accuracy and Consistency: Machine learning algorithms can analyze large amounts of data and identify complex patterns that may not be easily detected by human experts. This can lead to more accurate and consistent assessments of wine quality. By relying on objective data-driven models, the subjectivity and variability inherent in human evaluations can be reduced.

2. Efficiency and Cost-Effectiveness: Traditional methods of wine quality testing often involve manual labor and extensive sensory evaluations by experts. Machine learning can automate the process, allowing for faster and more efficient assessments. This can save time and resources, especially when dealing with large volumes of wine samples.

3. Scalability and Adaptability: Machine learning models can be trained on large datasets and can handle a wide range of wine samples. This scalability and adaptability make it easier to analyze different wine varieties, regions, and vintages. The models can also be updated and retrained as new data becomes available, ensuring continuous improvement and relevance.

4. Enhanced Decision-Making: Machine learning can provide valuable insights into the factors that contribute to wine quality. By analyzing the relationships between various parameters and quality ratings, winemakers can make informed decisions to optimize their production processes, improve quality control, and enhance the overall quality of their wines.

5. Consumer Preferences: Machine learning models can help identify patterns in consumer preferences by analyzing data from wine ratings, reviews, and sales. This information can assist winemakers in understanding consumer trends and tailoring their products to meet market demands.

6. Innovation and Experimentation: Machine learning can enable winemakers to explore new possibilities and experiment with different techniques. By analyzing the relationships between production practices, grape characteristics, and wine quality, machine learning models can provide insights that drive innovation in winemaking processes.

7. Brand Differentiation: Accurate and reliable wine quality testing using machine learning can help wineries establish a reputation for producing high-quality wines. By consistently delivering wines of superior quality, wineries can differentiate themselves from competitors and build a loyal customer base.

Overall, the motivation for wine quality testing using machine learning stems from the desire to improve accuracy, efficiency, and decision-making in winemaking processes, while also meeting consumer expectations and driving innovation in the industry.

1. RELATED WORK

Innovations in data mining have been used to plan wine quality. The goal of machine learning approaches, such as those used in a variety of applications, is to create models using data to predict wine quality. Since there is a lot of data in data mining and complexity is rising quickly, knowledge extraction from raw data is becoming more valuable and significant. Regression, classification, clustering, frequent item set mining, neural networks, and genetic algorithms are a few of the crucial techniques. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples.

It is given as

𝐴ccuracy = Number of Correct Predictions / Total number of predictions Made.

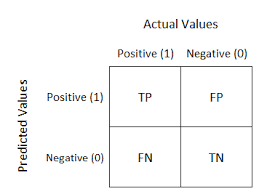


Fig. 1. Confusion Matrix

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model.

Where, TP: True Positive FP: False Positive FN: False Negative TN: True Negative.

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal.

It is given as-

𝐴ccuracy = (TP+FN)/N, Where, N:Total number of samples

F1 score-It is used to measure a test’s accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

𝐹1 = 2 ∗ 1/ (( 1/precision) + ( 1/recall))

F1 Score tries to find the balance between precision and recall.

1. PROPOSED METHODOLOGY

To conduct the study, the dataset was taken from the UCI Machine Learning Repository. The dataset for wine data has 6498 instances and 12 variables. Based on the inputs used, the data evaluation is completed with a prediction of the wine quality. In this dataset, attributes are predicted, and those that are less than or equal to '6' indicate bad wine quality and those that are greater than '6' excellent wine quality.

Type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, and alcohol represent a few of the highlights. The wine's acidity and basicity are represented by the pH value [3]. Wines that can be consumed have a pH scale between 3 and 4. The salt content represents the wine's chloride content. Using a range of physicochemical qualities, such as acidity and liquor properties, the information file aims to predict the rating that the master will accommodate a wine test [4-7].

The attributes used in the algorithm are shown in the Figure 2.

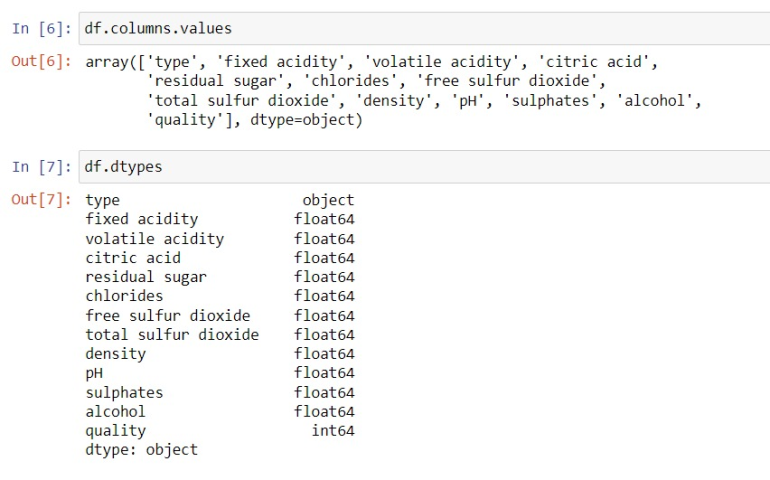


Fig. 2. Attributes of the dataset

The first step of the KDD (knowledge discovery for databases) process comes after data selection: data pre-processing. Datasets must be pre-processed in accordance with the specifications of the model being used before being inserted into it [8]. Eighty percent of the dataset is used as the training set, while twenty percent is used for testing. Training data are used to make the model match the data the best, and test data are used to assess the model's effectiveness.

For improved prediction, the data needs to be normalised or standardised because some attributes' values are simply too high and may overpower those of other attributes. Standardisation is done to put all attributes on an even playing field. Standard deviation and mean are also included.

1. ALGORITHMS USED

*A. Logistic Regression*

Logistic regression is a classification model, not a regression model. Observations are divided into a finite number of classes by this algorithm. In contrast to linear regression, which yields continuous numeric values, logistic regression modifies its output using the logistic sigmoid function to produce a probability value that may subsequently be translated to two or more discrete classes. This classification model works brilliantly with linearly separable classes and is exceedingly easy to use. It is a categorization method that is commonly used in business. The logistic regression model is a technique for binary classification that can be extended to multiclass or multi attribute classification.

*B. Random Forest*

Random forest and other supervised machine learning algorithms are frequently used in classification and regression problems. It builds decision trees on different samples, uses their majority vote for classification, and uses their average in regression scenarios. One of the most important features of the Random Forest Algorithm is the capacity to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification. It produces better results when it comes to classification problems.

*C. Voting Classififer*

The voting classifier makes a final prediction by combining the predictions of various distinct classifiers. It uses the weighted average (soft voting) or majority vote (hard voting) of the predictions made by its individual classifiers.

Both binary and multi-class classification issues are addressed by it. By combining predictions from various classifiers, it taps into the collective wisdom of the community, improving overall performance and producing forecasts with greater sturdiness. given data point.

*D. Adaboost Classififer*

We continue to utilise the same model even though we may temporarily change the parameters or add more data. Each trained model must be utilised on its own set of data, even if we form an ensemble. Boosting uses a method that is more iterative. Even if it uses a more ingenious method, the process is still an ensemble one because all models are combined to perform the final one.

*E. Gradient Boosting Classifier*

Gradient boosting is a technique for ensemble learning that combines a number of weak classifiers (usually decision trees) to produce a powerful predictive model. By systematically including new models that correct the shortcomings of the old ensemble, it constructs the model in stages.

Each successive model in gradient boosting is fitted to the residual errors of the preceding models, which are the discrepancies between the predicted and actual values. This strategy enables later models to concentrate on examples that were challenging for the earlier models to anticipate. The results of all the weak classifiers' predictions are combined to get the final prediction. Classifiers that use gradient boosting have good prediction accuracy and are resistant to overfitting.

1. RESULTS AND ANALYSIS

Here, we present the individual graphs and confusion matrices for the machine learning algorithms that we utilised, allowing us to examine which method has created or predicted wine quality with the highest degree of accuracy.

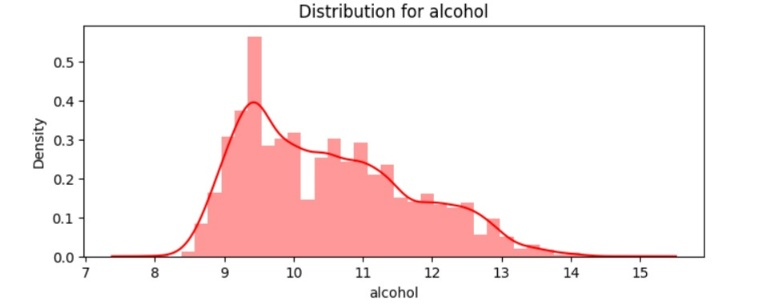


Fig. 3. Distribution for Alcohol

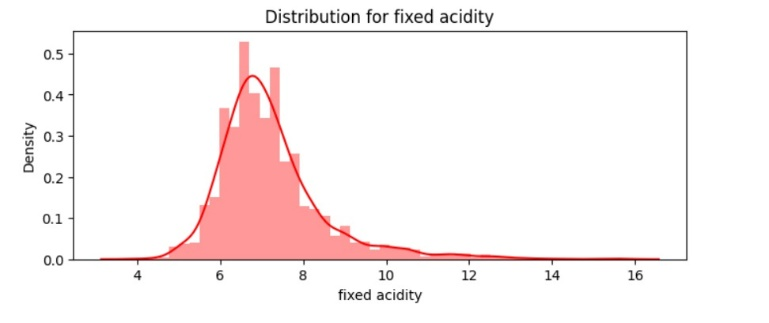


Fig. 4. Distribution for fixed acidity

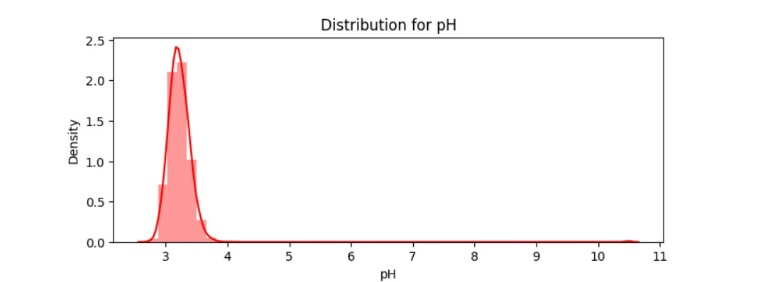


Fig. 5. Distribution for fixed acidity

1. *Logistic Regression*

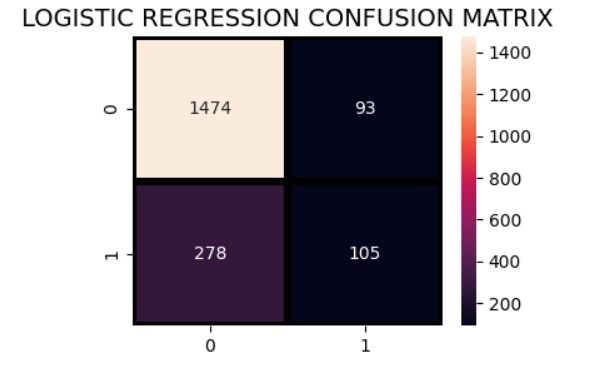


Fig. 6. Logistic Regression Confusion matrix.

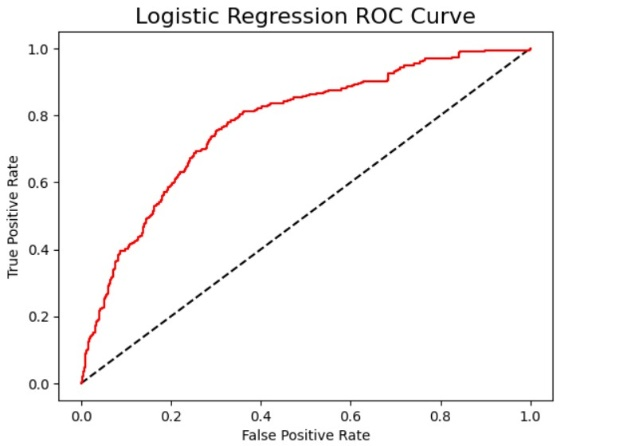


Fig. 7. Logistic Regression ROC curve.

1. *Random Forest*

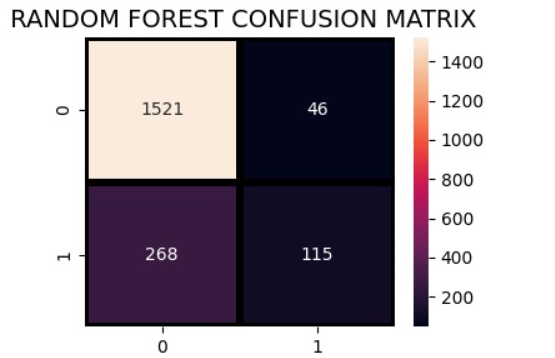
**

Fig. 8. Random Forest Confusion matrix.

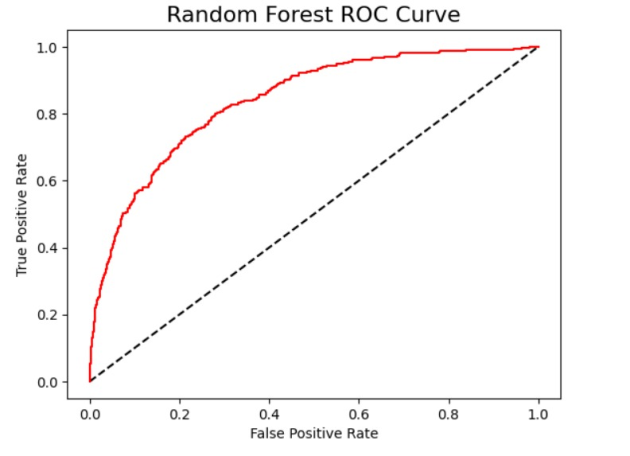


Fig. 9. Random Forest ROC curve.

1. *Voting Classifier*

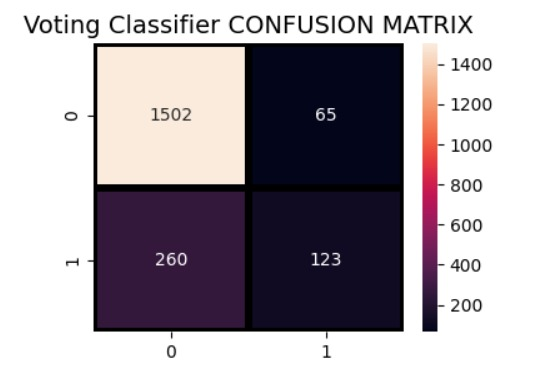


Fig. 10. Voting Classifier Confusion matrix.

1. *Adaboost Classifier*

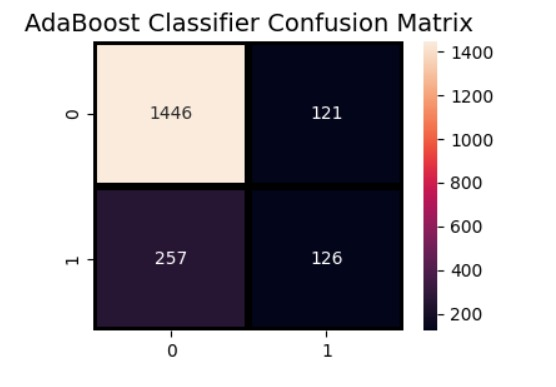


Fig. 11. Adaboost Classifier Confusion matrix.

1. *Gradient Boosting Classifier*

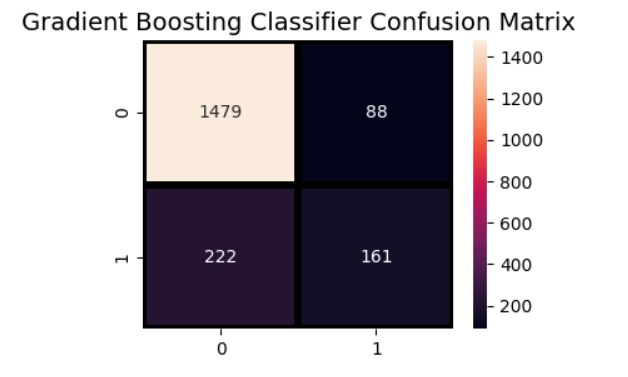


Fig. 12. Gradient Boosting Classifier Confusion matrix.

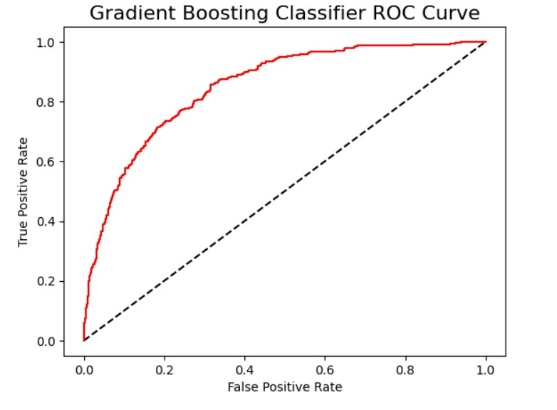


Fig. 13. Random Forest ROC curve.

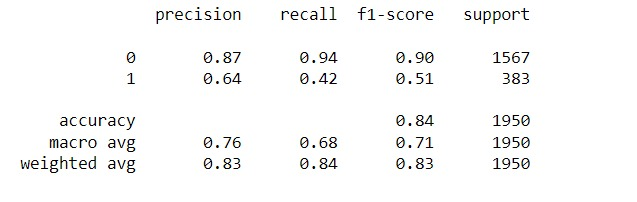


Fig. 14. Final Classification Report.

1. CONCLUSION AND FUTURE SCOPE

The dataset in this study is subjected to several machine learning algorithms, and classification is carried out using a variety of methods, with the Gradient Boosting Classifier achieving the highest accuracy, as seen in Figure 14. The experiment's dataset makes it clear that the Gradient Boosting Classifier improves wine quality prediction precision and accuracy.

The analysis of customer preferences and wine ratings can be included to this study. Machine learning models can provide consumers customized recommendations based on their taste preferences and prior experiences by merging this data with wine attributes.

1. REFERENCES
2. Lee, S., Park, J., & Kang, K. (2015). *Assessing wine quality using a decision tree. 2015 IEEE International Symposium on Systems Engineering (ISSE).* doi:10.1109/syseng.2015.7302752
3. A. Abdelhalim and I. Traore, “A new method for learning decision trees from rules,” IEEE International Conference on Machine Learning and Applications, pp. 693–698, Dec. 2009.
4. Kumar, S., Agrawal, K., & Mandan, N. (2020). *Red Wine Quality Prediction Using Machine Learning Techniques. 2020 International Conference on Computer Communication and Informatics (ICCCI).* doi:10.1109/iccci48352.2020.9104095
5. S. Ebeler, “Linking Flavour Chemistry to Sensory Analysis of Wine,” in Flavor Chemistry, Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409-422. [3] 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, 2016, pp. 137-151.
6. A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
7. 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, 73(4): 501-514, 2001.
8. 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., 57: 2421-2436, 2008.
9. Trivedi, A., & Sehrawat, R. (2018). Wine Quality Detection through Machine Learning Algorithms. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE). doi:10.1109/icrieece44171.2018.9009111
10. Sáenz-Navajas, M. P., Campo, E., Sutan, A., Ballester, J., & Valentin, D., “Perception of wine quality according to extrinsic cues: The case of Burgundy wine consumers,” Food Quality and Preference, 27(1), 2013, pp.44-53.
11. Gupta, Y., “Selection of important features and predicting wine quality using machine learning techniques,” Procedia Computer Science, 125, 2018, pp.305-312.
12. Freitas, A. A.: Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media, 2013.
13. Han, J., Pei, J., & Kamber, M.: Data mining: concepts and techniques. 3rd Edition. Elsevier, USA: Morgan Kaufmann, 2001.
14. Larose, D. T., & Larose, C. D.: Introduction to data mining and knowledge discovery. John Wiley & Sons, 1996.