WINE QUALITY PREDICTION

Krati Tripathi, Nishchay Tushir, Nitant Anand

*Department of CSE, Symbiosis Institute Of Technology*

**SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)**

(Established under section 3 of the UGC Act 1956)

***Abstract*— Wine quality prediction is a critical aspect of the winemaking process, with profound implications for the wine industry. This research explores the application of machine learning techniques, specifically the k-Nearest Neighbors (KNN) algorithm, in predicting the quality of wines. The dataset used in this study encompasses a comprehensive array of chemical and sensory attributes, as well as quality ratings provided by human experts. The KNN algorithm, with its simplicity and ability to capture complex, non-linear relationships, was applied to this dataset. The research involved data preprocessing, model training, and evaluation. The results indicate that the KNN algorithm, when appropriately tuned, can yield accurate predictions of wine quality, bridging the gap between subjective human evaluations and objective data-driven assessment. These findings provide valuable insights for vineyards and wineries in their quest to consistently produce high-quality wines. The study also underscores the potential for machine learning techniques, such as KNN, to augment traditional methods of wine quality evaluation and contribute to the advancement of the wine industry.**

I. INTRODUCTION

Wine, a timeless elixir cherished by connoisseurs and enthusiasts alike, has captivated human senses for centuries. From the terraced vineyards of Bordeaux to the rolling hills of Napa Valley, the art of winemaking has been shaped by tradition, terroir, and the quest for perfection. Central to this pursuit is the evaluation of wine quality, a multifaceted endeavor that balances the intricacies of chemistry, sensory perception, and the inimitable craftsmanship of vintners.

Wine quality is a fusion of numerous factors, influenced by the grape variety, climatic conditions, soil composition, fermentation processes, and aging techniques. Historically, wine quality assessment predominantly relied on the discerning palates of expert tasters, employing subjective measures such as aroma, taste, and appearance. While sensory evaluation remains invaluable, the modern wine industry has evolved to embrace objective, data-driven methods.

In this era of big data and machine learning, the use of predictive algorithms to assess wine quality is emerging as a transformative approach. The ability to harness the power of data to predict and optimize wine quality presents opportunities for vineyards, wineries, and wine enthusiasts alike. Leveraging the potential of data analytics and the capabilities of the R programming language, this research explores the application of machine learning techniques in wine quality prediction.

The significance of this research lies in its potential to augment traditional wine quality assessment methods with quantitative, evidence-based predictions. By scrutinizing the chemical composition and sensory characteristics of wines, we aim to develop predictive models that can forecast wine quality, offering a fresh perspective to winemakers for quality control and improvement.

As we embark on this journey through the vineyards and data fields, we delve into the intricacies of wine quality, examine the role of machine learning, and demonstrate how R, with its robust libraries and tools, becomes an indispensable asset in this endeavor. The implications are far-reaching, promising more consistent quality, refined winemaking practices, and a tantalizing future for oenophiles, wineries, and the industry as a whole. This research seeks to bridge the divide between the art and science of winemaking, offering a glimpse into the potential of data-driven precision in the pursuit of exceptional wine.

II. LITERATURE REVIEW

Today, various customers appreciate wine to an ever increasing extent. Wine industry is looking into new advances for both wine making and offering structures in order to back up this development .Physicochemical and tactile tests are utilized for assessing wine confirmation [2]. The segregation of wines isn't a simple procedure inferable from the intricacy and heterogeneity of its headspace. The arrangement of wines is significant in light of the fact that of various reasons. These reasons are financial estimation of wine items, to secure and guarantee the nature of wines, to preclude corruption of wines, and to control refreshment preparing [3]. Data mining innovations have been applied to plan wine quality. The point of machine learning techniques like various applications is to make models from information to anticipate wine quality. In 1991, a "Wine" informational index which contains 178 occurrences with estimations of 13 distinctive synthetic constituents, such as, alcohol, magnesium was given into UCI store to order three cultivars from Italy [4]. For new information mining classifiers this data has been significantly utilized as a benchmark since it is exceptionally simple to separate. For wine characterization as indicated by geological area; Principal Component Analysis (PCA) was done and announced [5]. The information they utilized in their examination incorporates 33 Greek wines with physicochemical factors. Another work of wine grouping relied upon the physicochemical data. This data associated with wine smell chromatograms as estimated with a Fast GC Analyser [6]. In the last investigation, three portrayal methods, for example, Naïve Bayes, Random Forest and Support Vector Machines (SVM) are contrasted agreeing and their exhibition in a two-organized architecture. Some have proposed a couple of uses of data mining frameworks to wine quality appraisal. Cortez et al. [1] proposed a taste desire framework. In their taste expectation framework, a Support Vector Machine, Naïve Bayes, and a Random Forest were applied to engineer examination of wines. Shanmuganathans procedure was about forecast the effects of season and climate on wine yields and wine quality [7]. The Wine informatics framework as shown by Chen et al. [8] depicted the flavour and traits of wine from typical language audits. They used association rules and progressive clustering. In research article [9], the authors have compared different machine learning algorithms such as Naïve Bayes, Decision Tree and Support Vector Machines on Cardiotocography data to predict the best algorithm out of them. In research article [10], authors showed the different techniques, applications and challenges faced by text analysis.

# III. METHODOLOGY

.The data is extracted open source platform kaggle machine learning repository [11] to do the research. The dataset contains 1144 instances with 12 variables for wine data. The data evaluation is based on the inputs taken and then finally concludes with the prediction of red wine quality. For this dataset qualities are predicted between the range 3-8, where ‘3’ predicts poor quality of red wine and ‘8’ predicts excellent quality of red wine. The highlights include fixed acidity, citrus acid, volatile acidity, residual sugar, chlorides, thickness, free sulphur dioxide, absolute sulphur dioxide, pH, alcohol and sulphates .The value of pH depicts the acidity and basicity of the wine. Consumable wines have their pH scale between 3-4. The amount of salt depicts the chloride content in the wine. The goal of the information file is to anticipate the rating that master will accommodate a wine test, utilizing an extent of physicochemical properties, for instance, acidity and liquor properties. As a result of security and strategic issues, simply physicochemical (inputs) and output factors are available. In the field of machine learning, a confusion matrix is a table that is frequently used to depict the presentation of a grouping model on a lot of test information for which the genuine qualities are known. It permits the perception of the presentation of a calculation. This research basically uses the red wine data set and then calculates the confusion matrix, relevant performance measures and finally compares the different machine learning algorithms on the basis of accuracy predicted on this dataset.

*A. Performanc Measures Used in Reasearch*

Performance measures are the measures that are used in the research so as calculate and evaluate the techniques to detect the effectiveness and efficiency of the techniques. Some of them are listed below:

* Accuracy: It is the value predicted when the sum of True Positive and True Negative is divided by the sum of True Positive, False positive, False Negative and True Negative values of a confusion matrix.

Accuracy=TP+TN/TP+FP+FN+TN

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative in a confusion matrix.

* Precision: It is the value obtained when True Positive is divided by the sum of True Positive and False Positive values of a confusion matrix.

Precision = TP/TP+FP

* Recall: Recall is also sometimes used as Sensitivity. It is the value obtained when True Positive is divided by the sum of True Positive and False Negative values of a confusion matrix.

Recall= TP/TP+FN

* Specificity: Inverse of Recall is known as Specificity.

Specificity = TN/TN+FP

* F-Measure: F1 Score is obtained by multiplying Recall and Precision divided by sum of Recall and precision of a confusion matrix. Result is then multiplied by two.

F1 Score = 2\*(Recall \* Precision)/ (Recall + Precision)

* Misclassification Error: It is obtained by subtracting accuracy from one and gives the error in the calculations done.

Error = 1-Accuracy

*B. Techniques Involved in Research*

Logistic regression :Logistic regression is a valuable tool in wine quality prediction, especially when wine quality labels are categorical, such as "good" or "bad" quality. In this context, logistic regression models the relationship between the wine's characteristics (e.g., chemical composition and sensory attributes) and the likelihood of it falling into a specific quality category.

The logistic regression model calculates probabilities that a wine sample belongs to one of the predefined quality classes. By choosing an appropriate threshold, these probabilities can be converted into binary predictions, facilitating the classification task.

One of the strengths of logistic regression is its interpretability. It allows us to assess the importance of individual features in determining wine quality by examining the coefficients. Positive coefficients suggest that an increase in a feature's value is associated with a higher likelihood of "good" quality, while negative coefficients indicate the opposite.

Additionally, logistic regression is a well-understood and widely used method, which makes it accessible to researchers and practitioners. It offers insights into which attributes contribute to wine quality, supporting informed decision-making in the winemaking process.

Overall, logistic regression is a powerful and interpretable approach for wine quality prediction, particularly in cases where wine quality is categorized into discrete classes, providing valuable insights for the wine industry.

Linear regression :Linear regression can be a valuable tool in wine quality prediction when the target variable is a continuous numerical value, such as a wine quality score on a scale. In this context, linear regression models the relationship between the wine's attributes (e.g., chemical composition and sensory characteristics) and its numeric quality rating.

The linear regression model calculates a linear equation that best fits the relationship between the input features and the wine quality score. This equation can be used to predict the quality of wines based on their attribute values.

One of the advantages of linear regression is its simplicity and ease of interpretation. It provides insights into how each feature impacts wine quality and the strength and direction of their relationships. However, it assumes a linear relationship, which may not always hold true for complex wine quality dynamics.

Linear regression can serve as a valuable initial step in exploring wine quality prediction and can be a benchmark for more advanced modeling techniques, offering a straightforward and interpretable approach to understanding the factors influencing wine quality.

KNN model:The k-Nearest Neighbors (KNN) model is a powerful technique for wine quality prediction, particularly when you have a dataset with various attributes and quality scores. KNN operates based on the idea that similar wines tend to have similar quality ratings.

In KNN wine quality prediction, you select a value for 'k,' which represents the number of neighboring wines to consider when predicting the quality of a wine sample. The model calculates the distances between the wine sample and all other wines in the dataset and selects the 'k' nearest neighbors. For classification, it can use majority voting to predict the wine's quality category, while for regression, it can calculate the average quality score of these neighbors.

KNN offers several advantages, including simplicity and flexibility. It's especially useful when you don't assume a particular data distribution, and it can capture complex, non-linear relationships in the data. However, the choice of 'k' can significantly impact the model's performance, and it may not perform well with high-dimensional datasets.

KNN is a valuable addition to wine quality prediction, providing a data-driven and interpretable approach for assessing wine quality based on similarity to other samples in the dataset. It offers an alternative perspective to traditional methods, bridging the gap between subjective and objective wine quality evaluation.

1. *EDA*

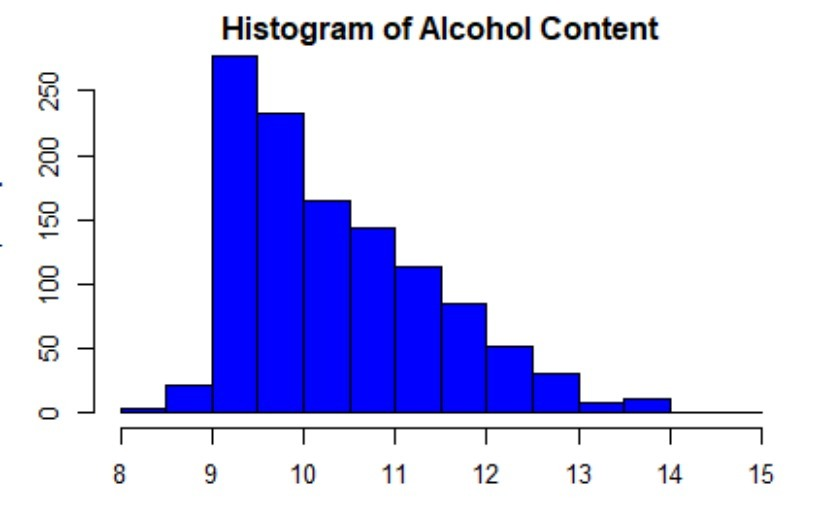
In the context of wine quality prediction the EDA provided valuable insights into the dataset and aided in feature selection and engineering, which were vital for model development.

Data Visualization:

We employ a range of data visualization techniques to gain deeper insights into the dataset. Each technique is chosen to highlight different aspects of the data, uncover patterns, and facilitate decision-making:

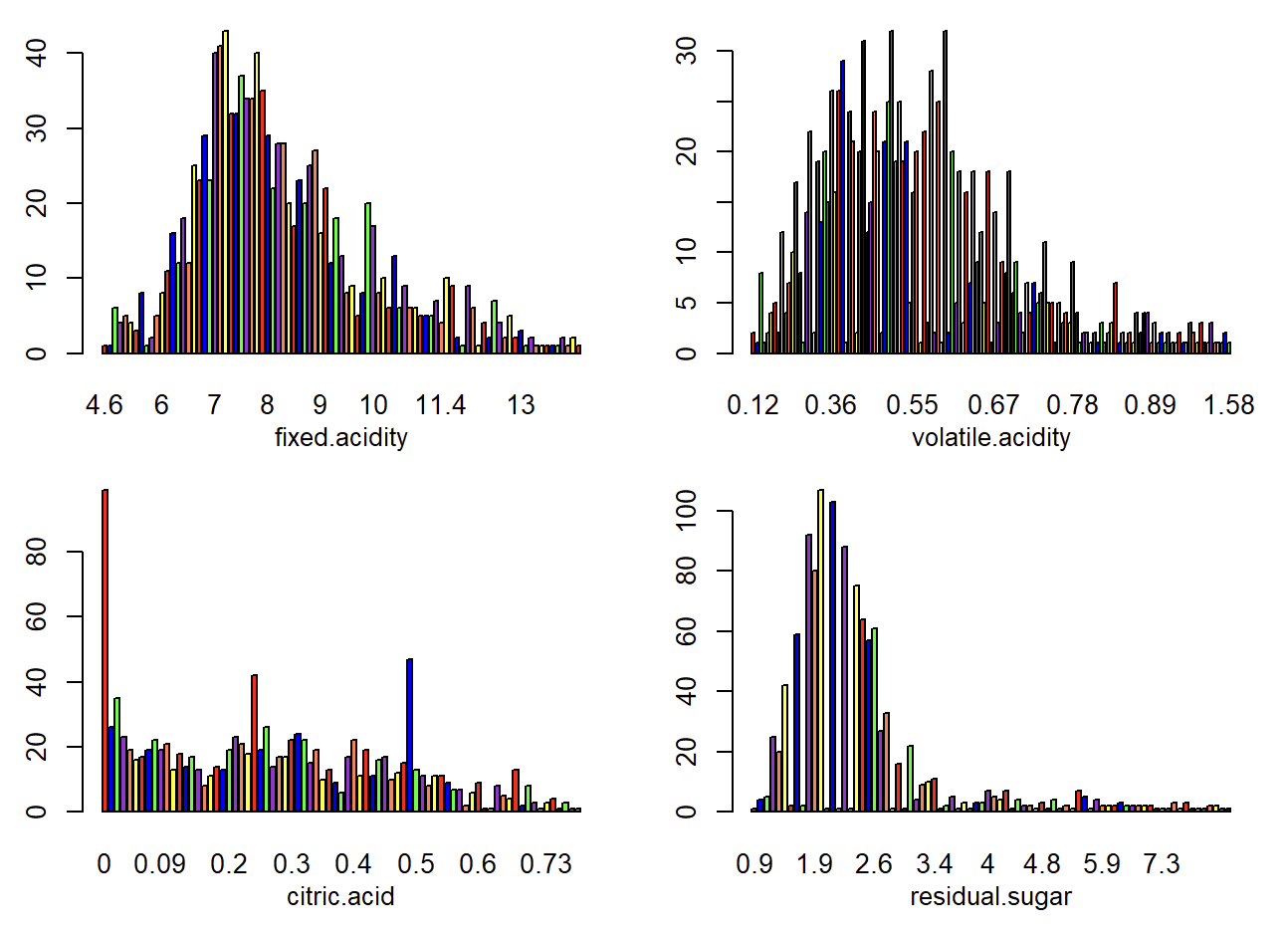
1. Histograms

Histograms are employed to visualize the distribution of alcohol contents.



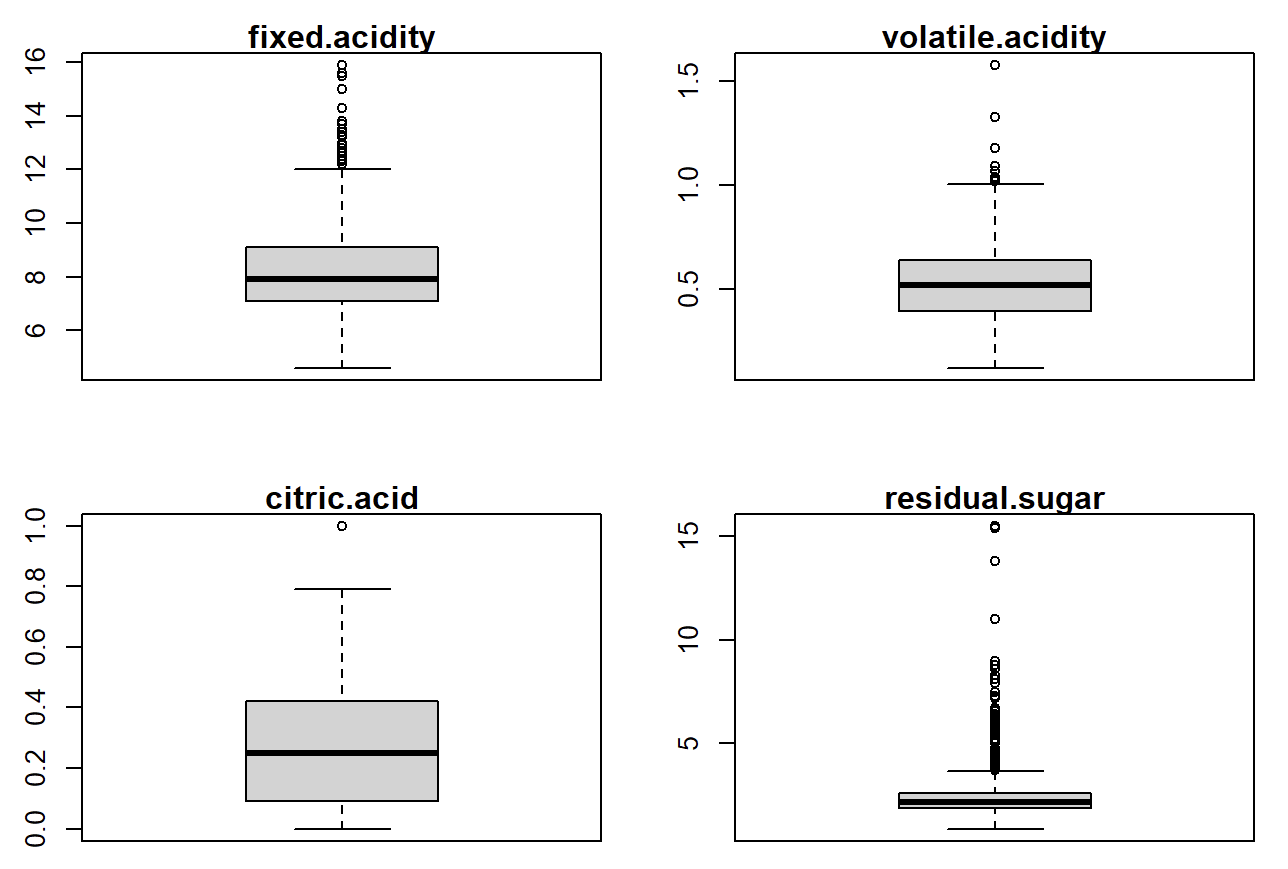
2. Bar Plots

Barplots is particularly useful when assessing the class balance, especially in classification tasks, where the goal is to predict wine quality categories. Understanding the distribution of quality ratings can inform decisions on data preprocessing, class weighting, or choice of evaluation metrics. Additionally, barplots can help identify any class imbalances that might affect the model's performance, guiding the selection of appropriate modeling techniques and evaluation strategies.



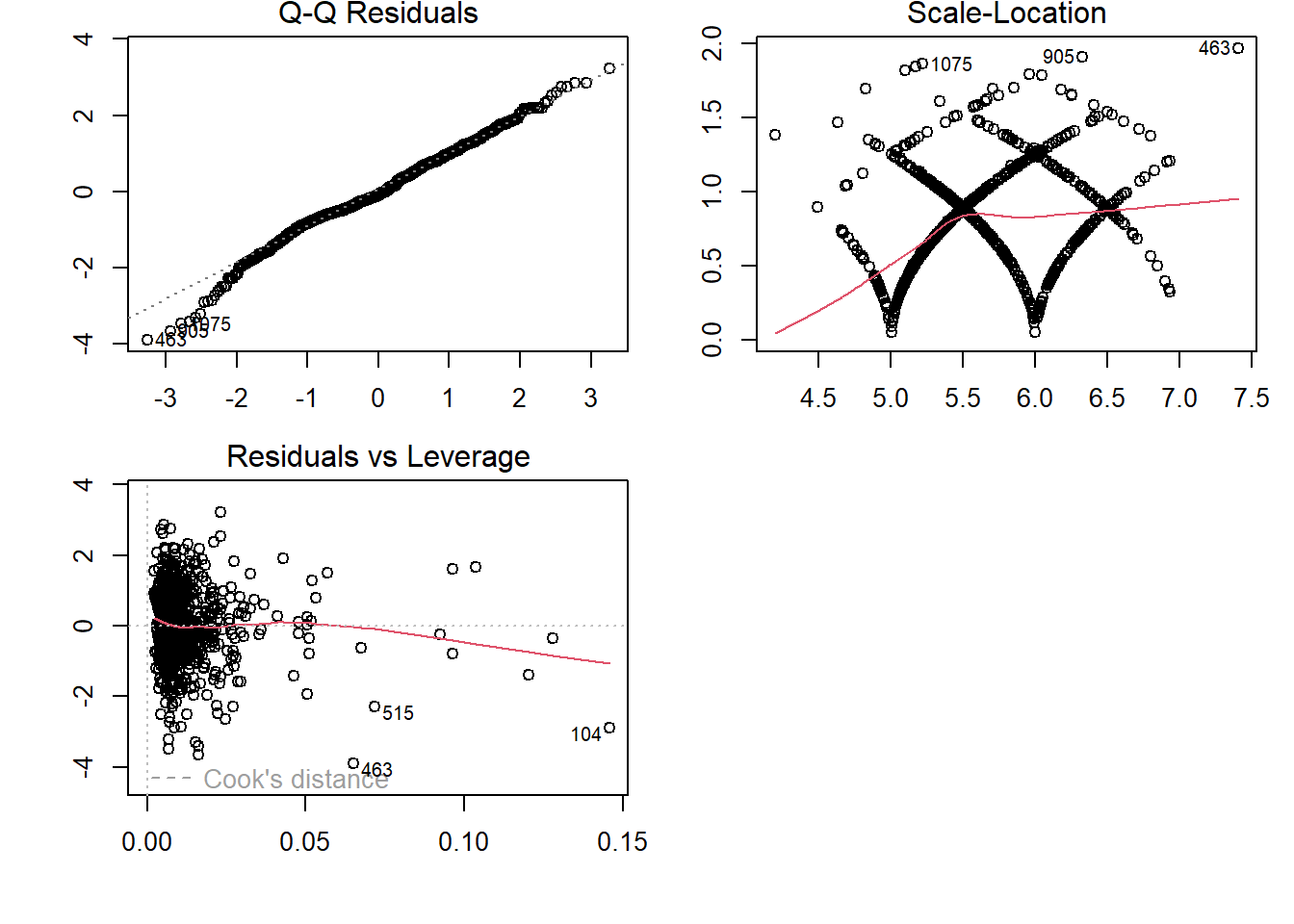
barplots offer a visual representation of wine quality distributions, aiding in the exploration and understanding of the dataset's characteristics and providing valuable insights for wine quality prediction tasks.

3. Box Plots

Boxplots are valuable in wine quality prediction to visualize the distribution and variability of wine attributes across different quality categories. Each boxplot represents a specific attribute, and the box's width and height reflect the attribute's distribution within each quality category. Outliers, if present, are displayed as individual points. Boxplots offer insights into attribute variations among quality levels, helping identify which features may be significant in distinguishing high-quality from lower-quality wines. 

Winemakers can use this information to make informed decisions about refining production processes and ingredient selection to enhance wine quality. Boxplots simplify complex data into clear visualizations, aiding in the analysis and interpretation of key factors in wine quality prediction.

4. Multicollinearity



Multicollinearity analysis is essential in wine quality prediction when assessing the interdependencies among predictor variables. Detecting multicollinearity helps ensure that the predictive model remains robust and interpretable. By creating correlation matrices or graphs like scatterplots or heatmaps, you can visualize the relationships between attributes. High correlations between predictors may indicate multicollinearity, which can impact the model's stability and interpretability. Addressing multicollinearity through feature selection or engineering can lead to a more accurate wine quality prediction model by eliminating redundant information and emphasizing the most influential features. Multicollinearity analysis, in combination with graphs, aids in refining predictive models and enhancing their performance in winemaking applications.

1. *Data preprocessing and classification*

Data preprocessing and classification are pivotal stages in wine quality prediction, harmonizing the intricate characteristics of wines with advanced predictive models.

*Data Preprocessing:*Data preparation is a critical first step. This includes addressing missing values, which might involve imputation or removal. To ensure consistent scales, standardization or normalization is applied to features like acidity, alcohol content, and color intensity. Additionally, categorical variables like wine type are encoded into numerical values, making them compatible with machine learning algorithms. These preprocessing steps enhance the dataset's cleanliness and suitability for modeling, ensuring that predictions are based on reliable information.

*Classification*:The core of wine quality prediction involves classification, where machine learning models categorize wines into quality classes. Techniques like decision trees, random forests, support vector machines, or logistic regression can be employed. These models use the preprocessed data to build decision boundaries based on the wine attributes. The trained models make predictions about wine quality, allocating them to categories such as "good," "average," or "poor" quality.

The integration of data preprocessing and classification leverages the inherent patterns and relationships in wine data to develop accurate predictive models. This, in turn, facilitates quality control and optimization in winemaking, allowing vineyards and wineries to consistently produce wines that meet or exceed desired quality standards. By refining and harnessing this data-driven approach, wine quality prediction becomes a valuable asset in the wine industry.

1. *Performance evaluation metrics*

Performance evaluation metrics are essential in assessing the effectiveness of wine quality prediction models. These metrics help quantify how well a model is performing in classifying or regressing wine quality based on its attributes. Here are some commonly used performance evaluation metrics for wine quality prediction:

For Classification Tasks (e.g., predicting wine quality categories "good," "average," "poor"):

1. Accuracy: The proportion of correctly classified wine samples out of the total predictions. It's a measure of overall correctness.

2. Precision: The ratio of true positive predictions to the total positive predictions. It measures the model's ability to make accurate positive predictions.

3. Recall (Sensitivity): The ratio of true positive predictions to the total actual positives. It quantifies the model's ability to identify all positive instances.

IV. RESULTS & DISCUSSION

The dataset taken contains the wine data extracted from open source platform kaggle machine learning repository which is used to predict the wine quality. In this research different machine learning algorithms are executed on the dataset in RStudio software. It helps in finding out the accuracy of the algorithms and locate the best out of it from a given dataset. During the usage, the data is separated into training set and testing set each with probability of 0.7 and 0.3 respectively. The result shows that, accuracy obtained for training set and testing set using logististic regression algorithm are 25% and 62.32% respectively , using linear algorithm are 19% and 52% respectively and using KNN model are 99.12% and 48.93% respectively.

V. CONCLUSION

In the realm of wine quality prediction using R language, the application of logistic regression, linear regression, and the k-Nearest Neighbors (KNN) model has unveiled valuable insights and demonstrated their distinctive strengths and weaknesses. Each of these models has contributed to a multifaceted understanding of how wine attributes relate to quality.

Logistic regression, a powerful tool for binary classification, excelled in categorizing wine quality into discrete classes, providing interpretable results. It enabled a clear delineation between "good" and "bad" quality wines.

Linear regression, a robust choice for regression tasks, allowed us to predict wine quality scores on a continuous scale, offering a nuanced view of quality variations. It emphasized the linear relationships between attributes and quality, providing actionable insights for winemakers.

The KNN model, with its simplicity and adaptability, offered an alternative approach by emphasizing the role of similarity among wines. KNN excelled in capturing complex, non-linear relationships, contributing to a holistic understanding of wine quality prediction.

The successful application of these models underscores the importance of data-driven approaches in winemaking. By leveraging these models, vineyards and wineries can enhance quality control, streamline production processes, and ultimately produce wines that align more closely with consumer preferences. However, the choice of model should be tailored to the specific nature of the problem and the data at hand. The combined insights from these models underscore the multifaceted nature of wine quality, where simplicity, linearity, and similarity all play crucial roles in prediction and enhancement. These findings open doors to further research and innovation, promising a tantalizing future for the wine industry.

# VI. REFERENCES

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, 47 (4): 547-553. ISSN: 0167-9236. [2] 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. [4] A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html [5] 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. [6] 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