IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Wine Quality Prediction Using Machine Learning

Mohan VH [1], Ramisetty Jyoshna [2], Girish Kumar KS [3], Gagan M [4] ,Nalluri Kavyasree [5], Prof. Shobana T S[6]
Assistant Professor
Department of Information Science and Engineering, B.M.S. College of Engineering

Abstract:

Predicting wine quality has emerged as a crucial topic of interest for producers and consumers alike, with the goal of ensuring production excellence and increasing market value. Relying on human expertise, traditional techniques of evaluating quality are time-consuming, subjective, and resource-intensive. Using publicly accessible datasets from the UCI Machine Learning Repository, we use Machine Learning (ML) approaches to forecast wine quality based on physicochemical properties in order to overcome these difficulties. Models like Logistic Regression, Decision Tree, Random Forest, and XGBoost are all part of our methodology and are assessed based on performance indicators like accuracy. By identifying the most pertinent qualities, feature selection reduces model complexity and boosts efficiency. Predicting wine quality has emerged as a crucial field of With XGBoost producing better results, our investigation demonstrates the strong predictive ability of ensemble methods and shows how they might be used to expedite quality assurance procedures in the wine business. This study demonstrates how ML-driven approaches can revolutionize conventional methods by providing a quicker, more accurate, and more affordable substitute for predicting wine quality. Predicting the quality of wine has become a crucial aspect of Our research demonstrates the strong predictive ability of ensemble approaches, with XGBoost producing better outcomes and showing promise for streamlining wine industry quality assurance procedures. This research highlights how machine learning (ML)-driven approaches can revolutionize conventional methods by providing a quicker, more accurate, and more affordable substitute for wine quality assessment.

Keywords: Wine Quality Prediction, Machine Learning (ML), Logistic Regression, Decision Tree, Random Forest, XGBoost, Feature Selection

1. Introduction

More than just a beverage, wine is a cultural staple and an art form that is valued for its unique Flavors and aromas everywhere in the world. Wine must be of high quality to be accepted by consumers and thrive in the competitive market. The market worth and reputation of wine are significantly impacted by its quality, which affects both casual and connoisseur consumers. Producers must maintain consistent quality, especially in a market where product certification and quality assurance play a major role in determining sales and customer trust.

In the past, wine quality evaluation was done after the fact, mostly depending on human specialists to judge sensory aspects like flavor, texture, and perfume. Although somewhat successful, this strategy had many drawbacks. Due to the vast range of individual quality assessments, sensory evaluations are by their very nature subjective. Furthermore, post-production quality inspections are expensive, time-consuming, and labor-intensive. Manufacturers frequently had to redo entire production processes if defects were found at this point, which resulted in significant financial losses.

Technology advancements have transformed the wine industry by allowing producers to integrate data-driven methods into their quality assurance procedures. During the actual production process, winemakers may forecast and maximize wine quality by gathering and evaluating data on physicochemical characteristics including acidity, sugar content, pH levels, and alcohol concentration. Openly accessible datasets, like those from Kaggle and the UCI Machine Learning Repository, have accelerated this change by providing structured information on red and white wine varieties with characteristics gleaned from in-depth physicochemical and sensory analyses.

Machine Learning (ML) has changed the game in a number of industries, including wine production, in recent years. ML approaches offer strong instruments for deciphering intricate datasets, identifying trends, and producing accurate forecasts. These techniques not only provide highly accurate wine quality predictions, but they also assist in determining the key determinants of wine quality. A crucial preprocessing step in machine learning, feature selection enables us to identify the most pertinent parameters while removing unnecessary ones, simplifying models, cutting down on training time, and improving overall performance.

This study investigates the capability of many machine learning algorithms, such as XGBoost, Random Forest, Decision Tree, and Logistic Regression, in predicting wine quality based on physicochemical characteristics. Our goal is to identify the factors that most affect quality by utilizing sophisticated feature selection techniques, giving winemakers useful information. Adopting such data-driven strategies reduces the need for subjective evaluations while enabling producers to optimize production procedures, leading to reliable and creative product offerings.

These prediction models also provide producers with additional experimental opportunities by enabling them to experiment with different parameter combinations to achieve unique wine profiles. In a market that is extremely competitive, this not only makes their items more distinctive but also boosts the value of their brand. Manufacturers can save time and money while continuously satisfying customer expectations by incorporating machine learning (ML) into the wine-making process and moving from reactive quality control to proactive quality assurance.

Our work demonstrates the revolutionary potential of machine learning (ML) in redefining conventional wine quality assessment methodologies in this age of rapid technological innovation. We hope to create a strong foundation for more inventive, economical, and efficient production methods by fusing data science and winemaking, thereby establishing new standards in the wine sector [1][2][3].

2. Data Description and Pre-processing

2.1 Overview of datasets

Two wine quality datasets, each representing a distinct variety of the Portuguese "Vinho Verde" wine, are used in this study. Samples of red and white wines are included in the datasets. There are 4,898 samples in the white wine dataset and 1,599 samples in the red wine dataset. Eleven physicochemical characteristics that characterize different aspects of the wine and one target variable that denotes quality are included in both datasets.

Acidity, sugar content, sulphur dioxide levels, pH, and alcohol % are examples of physicochemical characteristics. An ordinal scale is used to score the goal variable, wine quality, which represents tasters' subjective assessments. These datasets include a wide variety of inputs and a well-defined output variable, making them ideal for statistical and machine-learning analysis.

Fixed Acidity	Fixed acids such as tartaric acid in g/dm³.
Volatile Acidity	Acetic acid content in g/dm³.
Citric Acid	Citric acid content in g/dm³.
Residual Sugar	Sugar left after fermentation in g/dm³.
Chlorides	Salt content in g/dm³.
Free Sulfur Dioxide	Free SO_2 in mg/dm³.
Total Sulfur Dioxide	Total SO_2 in mg/dm³.
Density	Wine density in g/cm³.
Ph	Acidity level (pH scale).
Sulphates	Sulfates content in g/dm³.
Alcohol	Alcohol content in % volume.
Quality	Wine quality score (target).

Table 1: Attribute description

The performance of machine learning algorithms can be greatly impacted by outliers since they can skew the training process and result in the creation of less accurate and inferior models. Consequently, one of the most important steps in data pre-processing is locating and dealing with outliers. A popular technique for identifying outliers is the boxplot, which shows the distribution of the data visually. It identifies the distribution's general shape and draws attention to any possible outliers.

Figure 1 displays boxplots for every feature. All variables, with the exception of alcohol, clearly show skewness or the presence of possible outliers based on these graphs. Although eliminating extreme numbers may be one method of dealing with outliers, this strategy ought to be taken into consideration only in cases when it is evident that the values are inaccurate readings. Since we don't yet have enough data to prove that these extreme results are measurement errors, we haven't ruled them out [4].

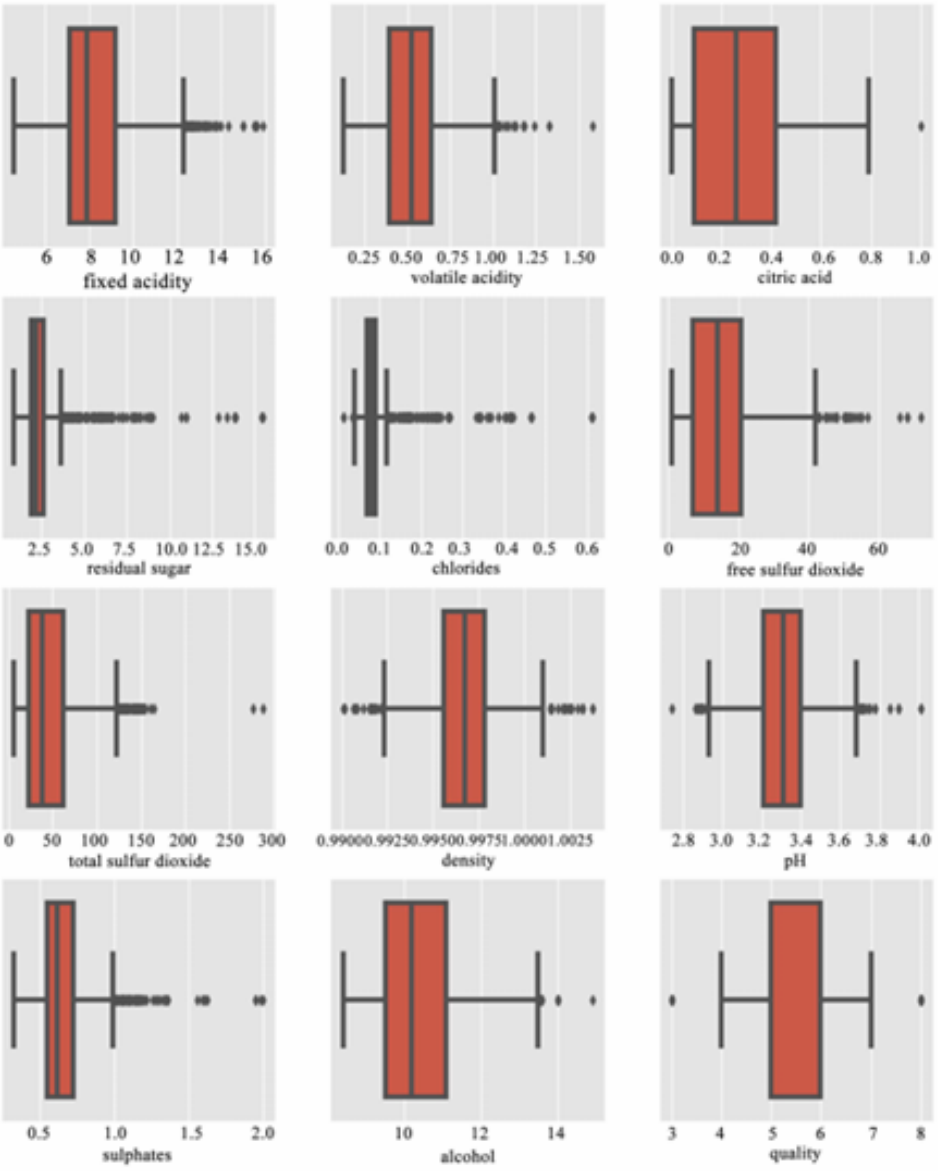

Figure 1: Box plot of the variables of the Redwine data.

3. Feature Selection

Feature selection is used to better comprehend the features and investigate how they relate to one another. The correlation between the features is determined using the Pearson coefficient correlation matrices.

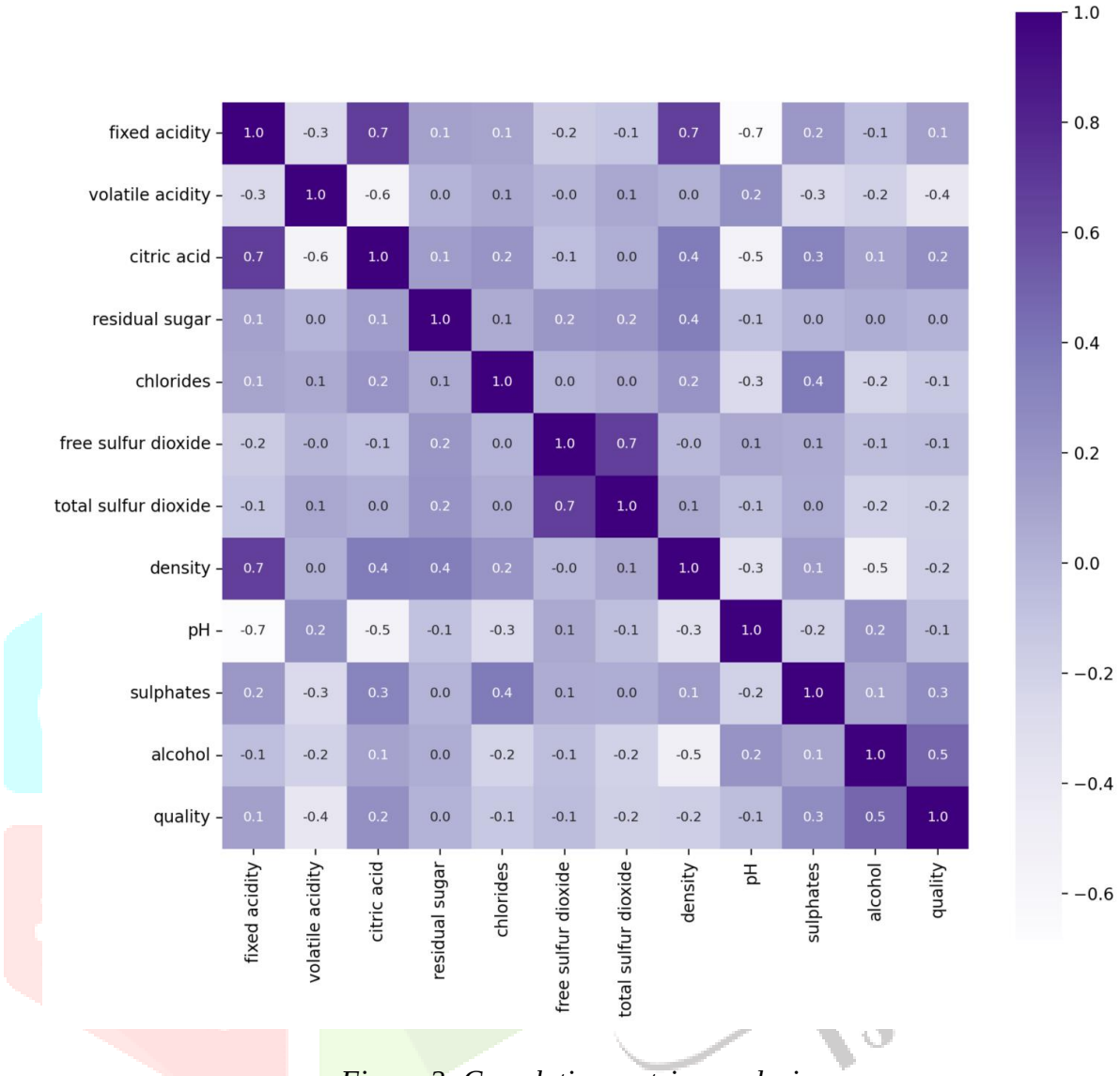

Figure 2: Correlation matrices red wine

The characteristics of red wine, such as "alcohol," "volatile acidity," "sulphates," "citric acid," "total sulphur dioxide," "density," "chlorides," "fixed acidity," "pH," "free sulphur dioxide," and "residual sugar," were ranked based on their high correlation values to the quality class in Figure 3.

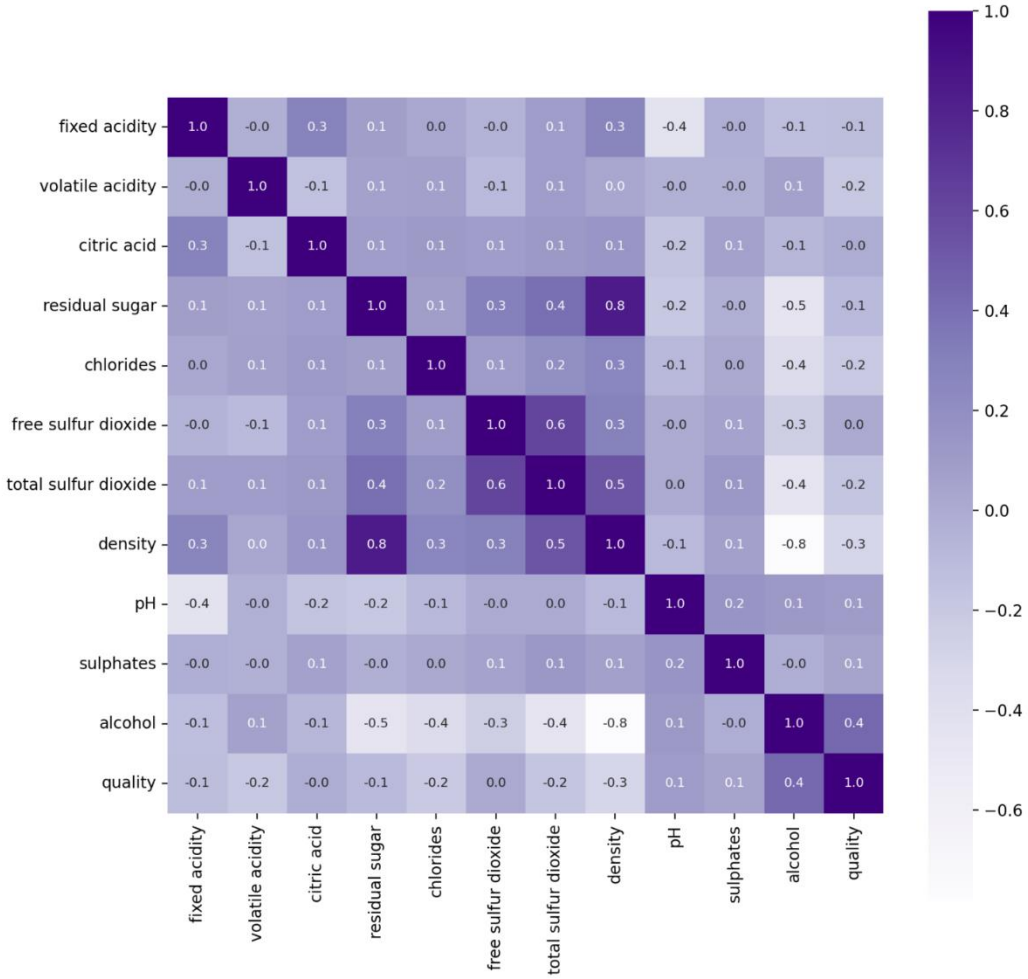

Figure 3: Correlation matrices white wine

The features "alcohol," "density," "chlorides," "volatile acidity," "total sulphur dioxide," "fixed acidity," "pH," "residual sugar," "sulphates," "citric acid," and "free sulphur dioxide" were also ranked based on its high correlation values to the quality class in Figure 4's white wine correlation matrix. [5][6].

4. Machine Learning Algorithms

To tackle different predictive demonstrating problems, machine learning provides a wide range of algorithms, each with distinct advantages and disadvantages. Out of all of them, supervised learning methods work well for categorization issues. In order to predict wine quality, we used a carefully chosen set of algorithms in this study: Random Forest Classifier, which improves accuracy through ensemble learning; Decision Tree Classifier, which is valued for its clarity in decision-making; Logistic Regression, which is known for its simplicity and interpretability; and the potent XGBClassifier, which is well-known for its effectiveness and reliable performance on structured data. These techniques were picked because they offer a thorough assessment of the prediction power of various algorithmic paradigms.

4.1 Logistic Regression

A popular statistical model for binary and multiclass classification issues is logistic regression. By employing the sigmoid function to represent the relationship between the dependent variable and the input features, it forecasts the likelihood of a particular outcome:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Here,

- P(y=1|X)P(y=1|X)P(y=1|X) is the probability of the target being 1 (e.g., higher wine quality).
- β0,β1,…,βn are the model coefficients.

A simple baseline model for dividing wine quality scores into predetermined groups is offered by logistic regression in the context of wine quality prediction. It may perform worse for complex, non-linear connections among physicochemical parameters, even while it performs well for linearly separable data.

4.2 Decision Tree Classifier

A non-parametric model called the Decision Tree Classifier uses straightforward decision rules to divide the data into subgroups according to feature values. A feature and a threshold are chosen at each node of the tree in order to minimize a criterion, like the Gini Index or entropy.

Gini Index is defined as:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

Where p_i is the proportion of samples belonging to class i, and C is the total number of classes.

Decision trees are excellent at capturing intricate relationships between characteristics (such as alcohol, acidity, and pH) in wine quality prediction. Since each path in the tree represents a series of choices that result in a specific quality score, they are simple to understand. They are susceptible to overfitting, though, particularly if the tree gets too deep.

4.3 Random Forest Classifier

Using random selections of data and characteristics, the **Random Forest Classifier** is an ensemble learning technique that creates numerous Decision Trees. It uses a majority vote to aggregate all of the trees' predictions (for categorization). This method improves robustness and decreases overfitting.

The model calculates the final prediction as:

$$\hat{y} = \text{Mode}(T_1(X), T_2(X), \ldots, T_k(X))$$

Where T_1, T_2,… , T_k are the predictions from k individual trees.

Random Forest is especially good at managing big datasets with noisy features when it comes to wine quality prediction. It captures intricate relationships and frequently performs better than stand-alone decision trees or more straightforward models like logistic regression.

4.4 XGBClassifier

A high-performance gradient boosting approach that produces models in a sequential fashion is called **XGBClassifier** (Extreme Gradient Boosting). Every new tree is built to fix the mistakes of the ones that came before it. The model uses gradient descent to optimize a loss function, like log-loss.

The objective function for XGBClassifier is given as:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k)$$

Where:

- $L(y_i, \hat{y}_i)$ is the loss function (e.g., squared error or log-loss).
- $\Omega(f_k)$ is the regularization term to prevent overfitting.

The XGBClassifier is well known for its capacity to process high-dimensional data and identify complex patterns. It is among the best-performing models for this job since it regularly produces high accuracy and good generalization performance when predicting wine quality.

Features of XGBClassifier:

1. Weighted Feature Handling: Features with stronger predictive power—such as alcohol content and volatile acidity in wine quality prediction—are given priority when XGBClassifier gives weights to them.
2. Built-In Regularization: L1 (Lasso) and L2 (Ridge) regularization are used into the technique to help minimize overfitting and guarantee that the model performs well when applied to unknown data.
3. Tree Pruning: It employs a technique known as Maximum Depth Pruning, which results in a more effective and straightforward model by halting tree growth when additional splits are ineffective.
4. Handling Missing Data: XGBClassifier can deal with missing data directly, eliminating the need for preparatory procedures like imputation.
5. Parallel Computing: Large datasets can benefit from Boost's ability to utilize parallel processing and its optimization for quick computation [6][7][8].

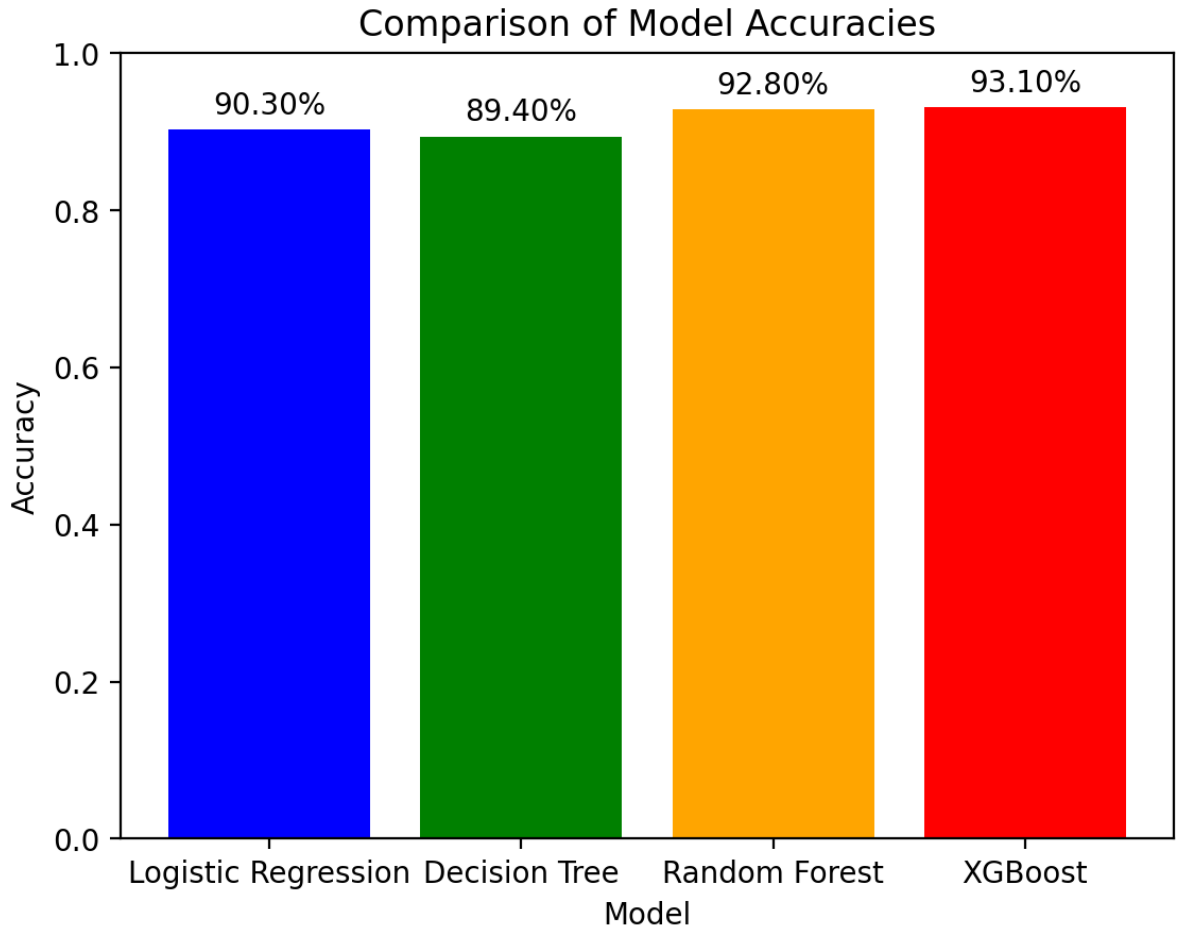

Figure 4: Comparison of Model Accuracies

5. Evaluation

Performance metrics are used to evaluate a model's efficacy and efficiency by calculating its predicted accuracy. There are four main categories into which predictions fall:

- True Positive (TP): Instances where the model correctly identifies a positive case.
- False Positive (FP): Cases where the model incorrectly predicts a positive result for a negative instance.
- False Negative (FN): Scenarios where the model fails to identify a positive instance and predicts it as negative.
- True Negative (TN): Instances where the model correctly predicts a negative case.

To evaluate the model's performance, the following method is applied:

1. Accuracy: The percentage of accurate forecasts among all predictions is represented by this metric. In order to calculate it, the total number of accurate predictions is divided by the entire number of evaluated occurrences.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative}}$$

2. Precision: Out of all the cases that the model predicts as positive, precision quantifies the percentage of real positive predictions. The number of anticipated positive cases that turn out to be true is the main focus.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3. Recall: The percentage of true positive cases that the model detects out of all actual positive cases is known as recall. Often called the True Positive Rate, it is computed as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

4. F1 Score: The harmonic mean of precision and recall is the F1 Score. It is used to evaluate a model's prediction accuracy, especially when dealing with unbalanced datasets. It is computed by multiplying the amount by two, dividing the precision and recall product by their sum.

$$\text{F1 score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Despite being a popular performance indicator, accuracy isn't always appropriate for datasets with unequal class distributions. Since accuracy tends to favour the dominant class in certain situations, it can be deceptive. As a result, the recall for minority classes is frequently 0, indicating that no occurrences of the minority class are detected by the model. In addition, the minority class's precision and recall are noticeably less than those of the majority class. Classifiers are challenged by this imbalance since they find it difficult to operate efficiently on the minority class while yet achieving high accuracy overall [6][9].

6. Conclusion

By displaying the effectiveness of models like Logistic Regression, Decision Tree, Random Forest, and XGBoost, this study shows the promise of machine learning approaches in predicting wine quality based on physicochemical attributes. Because of its excellent accuracy and resilience, XGBoost stood out among these as the most successful model, demonstrating its appropriateness for this field. Using feature selection techniques, we were able to pinpoint important characteristics that significantly affect wine quality, including sulphates, volatile acidity, and alcohol level. In addition to improving prediction accuracy, this method gives winemakers useful information that they may use to improve their production procedures and guarantee constant quality.

The study also emphasizes the shortcomings of accuracy as a metric for datasets that are unbalanced, stressing the significance of precision, recall, and F1 score for a thorough assessment. These results open the door for the wine industry to adopt data-driven approaches, which will allow producers to abandon the old, subjective method of evaluating quality in favour of a more efficient and objective one.

7. Future Work

While this study provides a solid foundation for wine quality prediction using machine learning, several avenues for future research can be explored:

1. Integration of Advanced Models**:** In order to capture intricate, non-linear correlations between features, future research may explore deep learning methods like neural networks.
2. Real-Time Monitoring**:** IoT sensors are being used in production facilities and vineyards to gather physicochemical data in real time for dynamic quality prediction.

3. Dataset Expansion**:** adding more datasets from various geographical areas and wine types in order to confirm and extrapolate the results.
4. Hybrid Techniques**:** creating hybrid models that further improve predicted efficiency and accuracy by combining the advantages of several methods.
5. Consumer Preferences: combining customer feedback and sensory data to match predictive models to the needs and tastes of the market.

Future studies can improve the use of machine learning in the wine business by tackling these issues, which will encourage creativity and guarantee higher-quality products in a variety of markets.

8.References

[1] Swapnil Darade, Nilesh Korade “Wine quality prediction” Volume 3, Issue 7 July 2021 and pp.: 1246-1252 www.ijaem.net.

[2] Prasanna M and Kamalesh Kumar. “Wine Quality Prediction using ML Techniques and KNIME” Volume 2, Issue 1, February 2022 ISSN (Online) 2581-9429.

[3] K.R. Dahal, J.N. Dahal, H. Banjade, S. Gaire. (2021). Prediction of Wine Quality Using Machine Learning Algorithms.

[4] A.C. Cortez et al., Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems.

[5] Prediction of Wine Quality Using Machine Learning Algorithms K. R. Dahal, J. N. Dahal, H. Banjade, S. Gaire.

[6] Wine quality prediction model using machine learning techniques by Rohan Dilip Kothawade

[7] Prediction of wine quality using machine learning algorithms open journal of statistics by k. R. Dahal et al. (2021)

[8] Machine Learning on Wine Quality: Prediction and Feature Importance Analysis researchgate

[9] https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

[10] Li, H., Zhang Z. and Liu, Z.J. (2017) Application of Artificial Neural Networks for Catalysis: A Review. Catalysts, 7, 306. https://doi.org/10.3390/catal7100306

[11] Shanmuganathan, S. (2016) Artificial Neural Network Modelling: An Introduction. In: Shanmuganathan, S. and Samarasinghe, S. (Eds.), Artificial Neural Network Modelling, Springer, Cham, 1-14. https://doi.org/10.1007/978-3-319-28495

[12] Jr, R.A., de Sousa, H.C., Malmegrim, R.R., dos Santos Jr., D.S., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira Jr., O.N. and Mattoso, L.H.C. (2004) Wine Classification by Taste Sensors Made from Ultra-Thin Films and Using Neural Networks. Sensors and Actuators B: Chemical, 98, 77-82. https://doi.org/10.1016/j.snb.2003.09.025

[13] 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, 547-553. https://doi.org/10.1016/j.dss.2009.05.016

[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, 451-464. https://doi.org/10.1002/sam.11474

[15] Lin, E.B., Abayomi, O., Dahal, K., Davis, P. and Mdziniso, N.C. (2016) Artifact Re moval for Physiological Signals via Wavelets. Eighth International Conference on Digital Image Processing, 10033, Article No. 1003355.

[16] Dahal, K.R. and Mohamed, A. (2020) Exact Distribution of Difference of Two Sam ple Proportions and Its Inferences. Open Journal of Statistics, 10, 363-374. https://doi.org/10.4236/ojs.2020.103024

[17] Dahal, K.R., Dahal, J.N., Goward, K.R. and Abayami, O. (2020) Analysis of the Res olution of Crime Using Predictive Modeling. Open Journal of Statistics, 10, 600- 610, https://doi.org/10.4236/ojs.2020.103036

[18] Crookston, N.L. and Finley, A.O. (2008) yaImpute: An R Package for kNN Imputa tion. Journal of Statistical Software, 23, 1-16. https://doi.org/10.18637/jss.v023.i10

[19] Dahal, K.R. and Gautam, Y. (2020) Argumentative Comparative Analysis of Ma chine Learning on Coronary Artery Disease. Open Journal of Statistics, 10, 694-705. https://doi.org/10.4236/ojs.2020.104043

[20] Caruana, R. and Niculescu-Mizil, A. (2006) An Empirical Comparison of Super vised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning, June 2006, 161-168. https://doi.org/10.1145/1143844.1143865

[21] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Sta tistical Learning: With Applications in R. Springer, Berlin, Germany.

[22] Joshi, R.P., Eickholt, J., Li, L., Fornari, M., Barone, V. and Peralta, J.E. (2019) Ma chine Learning the Voltage of Electrode Materials in Metal-Ion Batteries. Journal of Applied Materials, 11, 18494-18503. https://doi.org/10.1021/acsami.9b04933

[23] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Ma chine. Annals of Statistics, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451 [15] Chen, C.M. Liang, C.C. and Chu, C.P. (2020) Long-Term Travel Time Prediction Using Gradient Boosting. Journal of Intelligent Transportation Systems, 24, 109- 124. https://doi.org/10.1080/15472450.2018.1542304

[24] Turian, J.P., Bergstra, J. and Bengio, Y. (2009) Quadratic Features and Deep Archi tectures for Chunking. Human Language Technologies Conference of the North American Chapter of the Association of Computational Linguistics, Boulder, Colo rado, 31 May-5 June 2009, 245-248.

[25] Nwankpa, C., Ijomah, W., Gachagan, A. and Marshall, S. (2018) Activation Func tions: Comparison of trends in Practice and Research for Deep Learning. arXiv: 1811.03378

[26] Nair, V. and Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on International Con ference on Machine Learning, June 2010, 807-814.

[27] Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 15, 315-323.

[28] Amari, S. (1993) Backpropagation and Stochastic Gradient Descent Method. Neu rocomputing, 5, 185-196. https://doi.org/10.1016/0925-2312(93)90006-O

[29] Kingma, D.P. and Ba, J.L. (2014) Adam: A Method for Stochastic Optimization. ar Xiv:1412.6980

[30] Monro, T.M., Moore, R.L., Nguyen, M.C., Ebendorff-Heidepriem, H., Skourou mounis, G.K., Elsey, G.M. and Taylor, D.K. (2012) Sensing Free Sulphur Dioxide in Wine. Sensors, 12, 10759-10773. https://doi.org/10.3390/s120810759