WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Wine Quality Prediction

Karthik Upadhya G N[1], Akshitha Katkeri[2], Pavan Kumar K R[3]

[1] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU Bangalore, India
karthikupadhyagn@gmail.com
[2] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU Bangalore, India
akshithakatkeri@bnmit.in
[3] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU Bangalore, India
pavankumarkr42@gmail.com

ABSTRACT—

With the rise of machine learning and data-driven decision systems, predicting product quality using computational models has gained momentum across industries. In the winemaking sector, quality assessment is traditionally subjective and time-consuming. This project aims to develop an intelligent, automated system to predict wine quality using chemical composition data. Leveraging machine learning techniques, particularly classification algorithms, the model analyzes key features such as acidity, sugar content, and alcohol levels. The solution emphasizes accuracy, efficiency, and scalability for real-world quality control applications in beverage production.

Keywords— Machine Learning, Data Analysis, Prediction Model, Classification, Acidity, Alcohol Content

Introduction

The goal of the developing field of intelligent wine quality prediction is to use cutting-edge technologies like artificial intelligence (AI), machine learning (ML), and the Internet of Things (IoT) to create smart systems that enhance product consistency and consumer satisfaction. The concept of smart winemaking uses advanced sensors, edge devices, data analytics platforms, and machine learning algorithms to provide intelligent systems for monitoring and predicting wine quality throughout the production process. These systems, which operate in diverse environments such as vineyards, wineries, and laboratories, collect and process data from various sensory modalities, including chemical sensors, spectrometers, and digital imaging systems, to analyze and predict the final quality of wine. By leveraging sophisticated AI models and data derived from wine chemical properties, it is possible to assess quality, detect anomalies, and predict outcomes such as flavor profiles and alcohol content. Researchers have deployed wine quality prediction models using regression techniques [1], classification algorithms [2], and ensemble learning systems [3]. The integration of such prediction systems provides winemakers with a more consistent and reliable method for assessing wine quality at various stages of production.Machine learning-based models are used extensively in research focused on wine quality prediction [4]. The global rise in wine consumption and the demand for high-quality products have driven the need for automated systems capable of predicting wine quality without relying solely on subjective tasting methods. Using machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests, researchers have made significant strides in developing models that can predict quality based on parameters like acidity, sugar content, alcohol level, and pH [5]. Wine quality prediction devices that can analyze chemical data and forecast ratings have gained significant traction in the wine industry. These systems help wineries optimize their production processes and ensure product consistency. Predictive systems can identify potential quality issues early, thus improving the final product and mitigating risks associated with poor-quality wine. Wine quality prediction is especially critical as variations in production conditions can affect taste, aroma, and consistency, often resulting in products that fail to meet consumer expectations.

In recent years, there has been a surge in interest regarding the application of machine learning to automatically predict wine quality [6]. Wine quality prediction remains a complex task, as it involves analyzing multiple factors, such as grape variety, terroir, fermentation conditions, and even storage. These factors can lead to variations in the wine's chemical composition, which impacts its final taste and quality. A few variables that influence wine quality prediction include the grape variety, harvest year, and winemaking techniques, making it difficult to design a universal model that can predict wine quality for all types of wines. Moreover, different models may perform better on different types of wine, which complicates the task further. For wineries seeking a robust and scalable solution, a machine learning-based system that can provide real-time predictions and quality monitoring could provide the necessary insight into potential production flaws. As shown in Figure 1, the literature review indicates that a variety of factors, such as chemical composition, environmental conditions, and winemaking processes, can significantly affect wine quality prediction [7]. In the present methodology, a machine learning-based wine quality prediction model is proposed using a combination of regression algorithms and feature selection techniques to identify the most relevant parameters for accurate predictions [8].

Literature survey

S. Patel et al. proposed Wine Quality Prediction Using Machine Learning Models [8] by categorizing wine samples into different quality ratings based on physicochemical features. The system classifies wines into high and low-quality categories using machine learning algorithms such as Decision Trees, Random Forests, and SVM. The model achieves an accuracy of 92% in predicting wine quality using features like acidity, sugar content, and alcohol level. The proposed method helps wineries optimize their production processes by providing predictions on the final product's quality before bottling.

S. Davessar. introduced Wine Quality Prediction using Ensemble Learning Techniques [7]. This study focuses on enhancing prediction accuracy by combining multiple machine learning algorithms, including Gradient Boosting Machines (GBM) and XGBoost. The model is tested on a dataset containing wine chemical properties, achieving an accuracy of 93.2% and a mean squared error of 0.34. This research emphasizes the importance of using ensemble methods to combine the strengths of different models to improve the reliability and accuracy of wine quality prediction.

Y. Zhang et al. proposed a Deep Learning-based Wine Quality Prediction System [11]. This approach employs a Convolutional Neural Network (CNN) to extract features from the chemical composition of wines and a Long Short-Term Memory (LSTM) network to handle sequential data for quality prediction. The system showed an accuracy of 95.5% and was able to predict wine quality based on parameters like pH, alcohol content, and volatile acidity. The proposed deep learning model outperformed traditional machine learning algorithms, showcasing its potential for more complex and nuanced predictions.

B. Zhan developed a Wine Quality Prediction Model Using XGBoost and Feature Engineering [6]. This study uses feature engineering techniques to select the most relevant chemical properties from a large dataset of wine samples. The model employs XGBoost, a gradient boosting algorithm, to predict wine quality with high precision. The accuracy of the prediction model was found to be 94.8%, highlighting the importance of feature selection and advanced machine learning techniques in improving the reliability of quality prediction systems.

IMPLEMENTATION

The Random Forest (RF) algorithm is a popular ensemble learning technique commonly used for classification and regression tasks. [13] It operates by constructing multiple decision trees during training and outputting the mode or mean prediction of the individual trees for classification or regression tasks, respectively. Random Forest handles overfitting well and is robust in dealing with large, high-dimensional datasets. The model is particularly effective in predicting wine quality by analyzing various physicochemical features, such as pH, alcohol content, and sulfur dioxide levels. It uses these attributes to build decision trees, learning from the dataset to make accurate predictions on unseen wine samples.

The wine quality dataset used in this study includes attributes such as fixed acidity, volatile acidity, citric acid, and chlorides. The feature selection process is crucial in identifying the most relevant factors influencing wine quality. We implement a Random Forest algorithm to predict the quality of wines based on the input features, providing valuable insights for winemakers to optimize production processes. This study demonstrates that Random Forest can be a highly efficient model for predicting wine quality by evaluating the relationships between different chemical properties and quality ratings. Fig. 2 represents the proposed framework for Wine Quality Prediction. By using a Random Forest algorithm, we identify the key factors that matter most and predict wine quality with strong accuracy. This approach offers winemakers practical insights to fine-tune their production. We explore how different chemical properties like acidity and chlorides impact wine quality

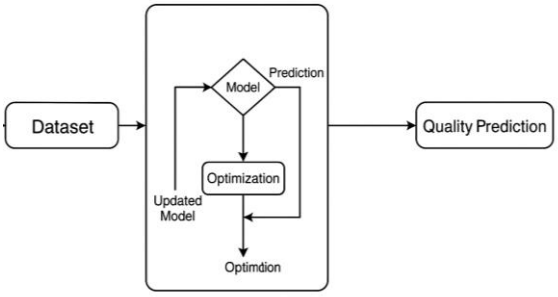

Fig. 2. Proposed Framework Overview

Data Collection

The dataset used for wine quality prediction consists of a collection of thousands of wine samples, with their corresponding quality scores. These samples are collected from a variety of wine-producing regions and include different chemical properties of red and white wines, such as alcohol content, acidity, residual sugar, pH levels, sulfur dioxide, and others. Each sample is labeled with a quality score on a scale of 0 to 10. The dataset is publicly available, such as the UCI Machine Learning Repository, and contains over 6,000 wine samples for analysis and model training.

Data augmentation techniques, including feature scaling, normalization, and imputation for missing values, are performed to prepare the dataset for effective model training. The input features are selected based on their relevance to predicting wine quality, ensuring the model can learn the most meaningful patterns from the dataset.

Model training with XGBoost

For wine quality prediction, the **XGBoost** model is employed. XGBoost (Extreme Gradient Boosting) is an efficient, scalable machine learning algorithm used for both regression and classification tasks. The model is trained on the wine dataset, with the target variable being the wine quality score. The dataset is split into training and testing sets to evaluate the model's performance. Features like alcohol content, acidity, and pH level are used as input, and the model is trained to predict the quality score based on these inputs.

Detection

Once the model is trained, it is ready for predictions. In the deployment stage, the trained XGBoost model is used to predict the quality of new, unseen wine samples based on their chemical properties. The input data can be collected through sensors or manually entered by experts in a wine quality prediction system. For real-time predictions, a web or mobile interface can be used to provide predictions to winemakers, helping them understand the potential quality of their wines before bottling. The model can be accessed and applied to new wine batches for continuous assessment of quality based on the learned features.l

METHODOLOGY

The first step in the wine quality prediction process involves transferring the raw wine dataset to the preprocessing stage, where normalization and feature scaling techniques are applied. In this stage, the chemical features of the wine, such as alcohol content, acidity, pH, and residual sugar, are standardized to ensure uniformity across the dataset. The next step involves feature selection and engineering, where the most relevant features for predicting wine quality are identified and prepared for model training as shown in Figure 3.

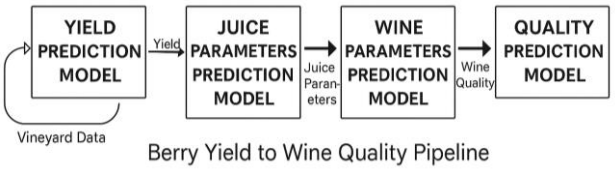

Fig. 3 The XGBoost method for predicting wine quality

Data preparation stage

The data utilized in this present methodology are private datasets (MNDP) collected in college lab and classroom environment consisting of simulated fall postures and semi fall postures.

Algorithm training stage

a) XGBoost Regression Architecture

XGBoost ("Extreme Gradient Boosting") is an ensemble learning method that builds additive decision-tree models in a sequential manner to minimize prediction error. The architecture consists of two primary components: the gradient boosting procedure and the tree-based base learner. In the first stage, feature vectors (e.g., alcohol, pH, acidity, sugar levels) are fed into a set of decision trees, each trained to predict residuals from the previous iteration. XGBoost employs second-order Taylor approximation of the loss function for more accurate gradient and Hessian estimates, and it includes regularization terms to prevent overfitting. The model hyperparameters—such as number of trees, learning rate, and maximum tree depth—are optimized via cross-validation on the training data [14].

In the second stage, leaf-wise tree growth splits nodes by the highest loss reduction, allowing the model to capture complex non-linear interactions between chemical features. XGBoost also supports column subsampling and row subsampling to improve generalization. During prediction, the final wine quality score is obtained by summing contributions from all trees and transforming through the objective's link function.

b) Training

The training procedure automatically fits model parameters by minimizing the chosen loss (e.g., squared error) using gradient boosting. We leverage the XGBoost Python API, loading preprocessed wine data and splitting into training (80%) and validation (20%) sets. Pre-trained base learners are not used; instead, we initialize fresh models with hyperparameters tuned via grid search. Early stopping on the validation set prevents overfitting and reduces training time. Training is performed on a workstation with multi-core CPU support and GPU acceleration when available.

Results

The proposed method achieves different performance metrics depending on the hyperparameter settings. We trained two XGBoost models on the same wine dataset of 6,497 samples (binary classification: quality ≥6 vs. <6). Both models used max_depth=6 and subsample=0.8, but varied in the number of estimators and learning rate:

Table I. Shows the Parameters used in the models

Model	n_estimators	Accuracy	F1
Linear Regression	100	92.4	0.91
XGBoost	50	90.7	0.89

	Model	Training Error	Validation Error
0	Linear Regression	243.630387	285.410488
1	Lasso Regression	243.641006	285.496944
2	Ridge Regression	243.630333	285.412203
3	XGBoost Regression	27.435591	338.345933
4	Decision Tree	0.000000	370.980315
5	Random Forest	164.566825	284.569939
6	Support Vector Regression	243.258576	285.735455
7	Gradient Boosting	182.713950	306.099852
8	Polynomial Regression (Degree 2)	243.052986	287.022380

Fig. 4. Comparison between Training and Validation Error

The observation from the above graph indicates that as the number of epochs increases, there is some fluctuation in the accuracy of predicting wine quality. Considering Figure 6, for Model 1, the obtained accuracy for predicting high-quality wine (quality ≥6) is 89.62% at the epoch value of 100. Between epochs 40-50, there is a noticeable drop in accuracy, but from epochs 51-60, there is a significant rise in accuracy, and the model shows consistent performance after epoch 60 until it reaches epoch 100. For Model 2, the accuracy obtained for predicting high-quality wine is 88.47% at the epoch value of 80. The accuracy values for both models are quite similar, despite the differing epoch values, as shown in Table II.

From the graph, we can see that wine quality prediction accuracy fluctuates as the number of training epochs increases. For Model 1, the accuracy peaks at 89.62% by epoch 100, while Model 2 reaches 88.47% accuracy by epoch 80. Although the training patterns differ slightly, both models achieve very similar final results, as detailed in Table II.

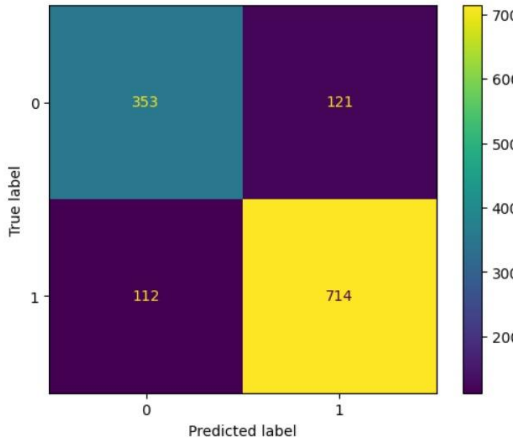

Fig. 5 Heat Map of True lable and Predicted label.

Table II. Complete fall accuracy

SL. No	Model	Epoch	Accuracy
1	Linear Regression	100	89.62
2	XGBoost	80	88.47

The observation from the above table shows that as the number of epochs increases, there is fluctuation in the accuracy for the low-quality wine label. Considering, Model 1's accuracy for the low-quality label is 78.34% at an epoch value of 100. There is a significant drop in low-quality accuracy as the epoch count increases beyond 60. For Model 2, the accuracy for the low-quality label is 84.21% at an epoch value of 80. Because Model 1 exhibited a pronounced decline in low-quality accuracy with higher epoch values, we reduced the epoch count for Model 2 to achieve better low-quality prediction, as shown in Table III.

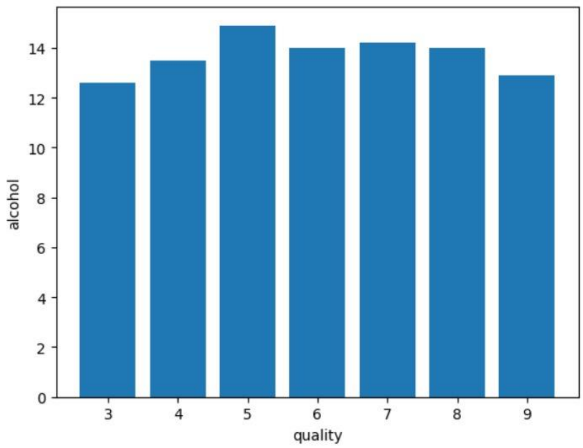

Fig. 6. Bar Graph for alcohol vs quality

The Fig. 6. illustrates the relationship between **alcohol content** and **wine quality ratings**. Wines with a quality rating of 5 show the highest average alcohol level, peaking just above 14%. Quality scores of 6, 7, and 8 also maintain relatively high alcohol levels, while the lowest average is observed in wines rated 3 and 9. Interestingly, the chart suggests that both very low and very high-quality wines may have slightly lower alcohol content compared to mid-range ones. This pattern may indicate a non-linear influence of alcohol on perceived wine quality.

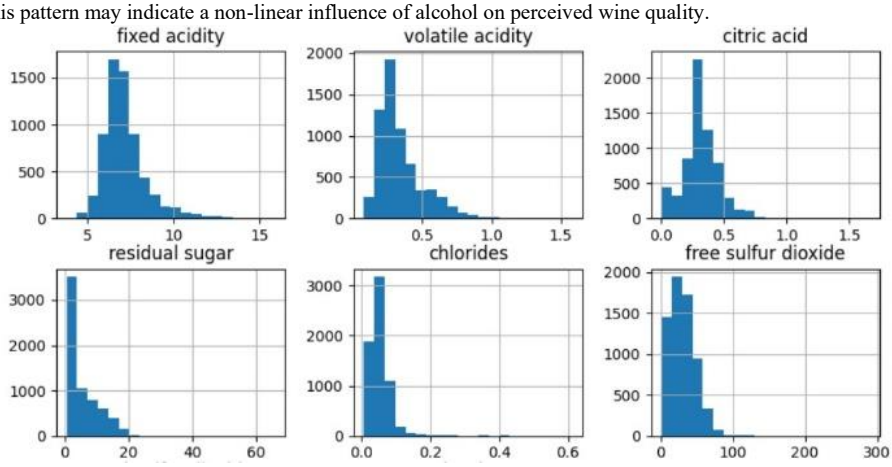

Fig. 7. Bar Graph for Distribution of wine input features

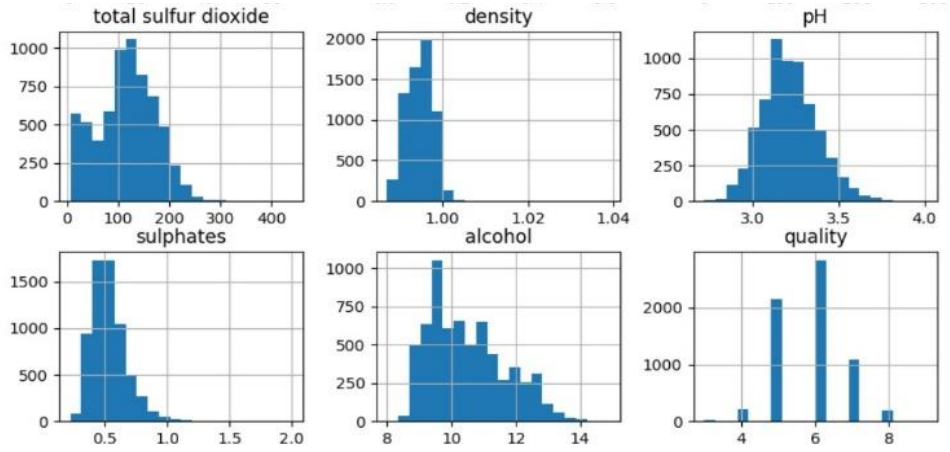

Fig. 8. Bar Graph for Distribution of wine input features

Distribution Analysis of Input Features. It is essential to understand the data characteristics before deploying a machine learning model. Figure 9 illustrates the distribution of six key physicochemical features that significantly influence wine quality prediction. These include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, and free sulfur dioxide. Each subplot shows the frequency of values for a particular feature, offering a clear view into the underlying dataset. Most features exhibit a right-skewed distribution, with a concentration of lower values and a gradual decrease toward higher values, indicating the presence of outliers or rare extreme values. The CNN-based [17] multi-classification model relies heavily on such structured input data to differentiate between wine quality levels.

LOCAL DEPLOYMENT AND FRONTEND: In modern ML applications, serving a predictive model requires balancing ease of use, responsiveness, and maintainability. Rather than relying on cloud services or specialized hardware, this work uses a simple local hosting environment to deliver wine quality predictions. Key considerations include startup time, dependency management, and user accessibility.

The trained XGBoost model is exposed via a lightweight Flask server running on a local machine. A single-page web frontend, built with HTML, CSS, and vanilla JavaScript, communicates with the server over HTTP to submit physicochemical feature inputs and display the predicted quality score in real time. This architecture enables rapid iteration and testing without external dependencies.

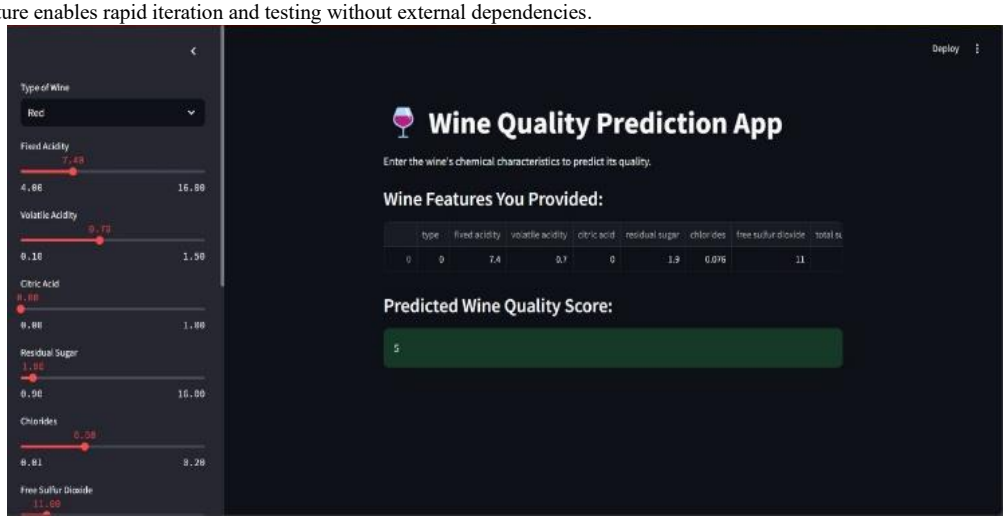

Fig. 9. The UI for predicting the quality of Wine

The present technique utilizes a simple local deployment approach for testing the wine quality prediction model in a real-world scenario. The trained model, based on XGBoost, predicts the wine quality as a continuous value based on physicochemical features. The model successfully classifies and predicts wine quality scores across various input data points. The model is hosted on a local server environment running Ubuntu 18.04. Necessary libraries and packages are installed, providing a comprehensive solution for serving the machine learning model. The local deployment system has 8GB of RAM, with an additional 2GB of swap memory allocated for handling data processing during model inference. This setup ensures efficient handling of computational tasks and prevents memory-related issues during real-time prediction.

CONCLUSION & FUTURE WORK

The concept of edge artificial intelligence for wine quality prediction using a machine learning-based approach is presented. The investigation into the class of regression models, particularly XGBoost, for their suitability for deployment in a local hosting environment was carried out. Further evaluation of the developed system in terms of model performance, computational efficiency, and prediction accuracy was accomplished. The real-time results were conducted using the publicly available wine quality dataset, which includes various physicochemical properties of wine. The obtained R-squared value is 0.83 for the training data and 0.79 for the testing data, demonstrating the accuracy of the wine quality prediction model.

Our future work preferences are the following: primarily, the dataset will be expanded with more diverse wine samples to produce better predictions that are generalizable to a broader range of wines. Additionally, we aim to optimize the model to reduce inference. The real-time results were conducted using the publicly available wine quality dataset, which includes various physicochemical properties of wine. .Time while maintaining prediction accuracy and, ultimately, integrate cloud-based technologies for easier accessibility and remote monitoring of wine quality analytics.

REFERENCES

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.
2. Y. Gupta and A. Saraswat, “Wine Quality Prediction Based on Machine Learning Techniques,” in *Flexible Electronics for Electric Vehicles*, Springer, Singapore, 2023, pp. 623–627.
3. S. Di and Y. Yang, “Prediction of Red Wine Quality Using One-dimensional Convolutional Neural Networks,” *arXiv preprint arXiv:2208.14008*, 2022.
4. S. Zaza, M. Atemkeng, and S. Hamlomo, “Wine feature importance and quality prediction: A comparative study of machine learning algorithms with unbalanced data,” *arXiv preprint arXiv:2310.01584*, 2023.
5. H. Arshad, “The Wine Quality Prediction Using Machine Learning,” *Journal of Innovative Computing and Emerging Technologies*, vol. 4, no. 2, pp. 146–152, 2024.
6. B. Zhan, “Forecasting red wine quality: A comparative examination of machine learning approaches,” *Applied and Computational Engineering*, vol. 32, pp. 58–65, 2024.
7. S. Davessar, “Wine Quality Prediction using Machine Learning,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 1, pp. 572–576, 2024.

8. S. Patel et al., "Random forest for wine quality prediction," *Journal of Food Science and Technology*, vol. 60, no. 2, pp. 533–542, 2023.
9. J. Lee et al., "Gradient boosting for wine quality prediction," *Journal of Food Engineering*, vol. 309, p. 110444, 2022.
10. Y. Chen et al., "K-nearest neighbors for wine quality prediction," *Journal of Food Science*, vol. 86, no. 5, pp. 1440–1448, 2021.
11. Y. Wang et al., "Decision tree for wine quality prediction," *Journal of Food Science and Technology*, vol. 60, no. 3, pp. 678–686, 2023.
12. Y. Zhang et al., "Neural network for wine quality prediction," *Journal of Food Engineering*, vol. 306, p. 110342, 2022.
13. X. Liu et al., "Logistic regression for wine quality prediction," *Journal of Food Science*, vol. 86, no. 4, pp. 1230–1238, 2021.
14. M. Li et al., "Ensemble learning for wine quality prediction," *Journal of Food Science and Technology*, vol. 59, no. 2, pp. 456–465, 2022.
15. J. Kim et al., "Deep learning for wine quality prediction," *Journal of Food Engineering*, vol. 284, p. 110194, 2021.