IJARSCT

ISSN (Online) 2581-9429

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 3, November 2024

Wine Quality Prediction using Machine Learning

Dr G Paavai Anand[1], Rithvika A D[2], Mangai Shreya[3], Sneha Vijaykumar[4]

Assistant Professor (SG), Department of CSE[1]

Students, Department of CSE(E-Tech)[2,3,4]

SRM Institute of Science and Technology, Vadapalani, Chennai, TN, India

Abstract*:* *This project investigates a basic machine learning approach to predicting wine quality based on chemical properties. Using a dataset containing attributes such as acidity, sugar content, pH, and alcohol level, the study classifies wines into quality categories. Simple machine learning algorithms, including logistic regression and decision trees, were employed to develop and evaluate predictive models. To enhance performance, data preprocessing and feature selection techniques were applied. The findings highlight alcohol and volatile acidity as key predictors of wine quality, though the model's accuracy reflects the limitations of the simplified approach. This project illustrates the potential of basic machine learning methods for preliminary wine quality assessment, with opportunities for improvement through advanced techniques and larger datasets.*

Keywords: Wine quality prediction, machine learning, SMOTE, Logistic Regression, Decision Tree, Random Forest, feature selection.

I. INTRODUCTION

Wine quality plays a crucial role in determining a product’s appeal and marketability, impacting both consumer trust and brand reputation. High-quality wine not only allows producers to command premium prices but also fosters loyalty through certifications that assure consistent standards. Traditional methods for assessing wine quality rely on expert tasters evaluating sensory characteristics like taste and aroma. While effective, these approaches are subjective, labor-intensive, and costly, with results often influenced by individual preferences or external factors. As global wine consumption rises, the industry faces the challenge of implementing efficient, objective, and scalable quality assessment processes to meet demand.

This project leverages machine learning to predict wine quality based on chemical attributes such as acidity, pH, and alcohol content, using a dataset sourced from Kaggle. By employing feature selection, the study identifies key properties influencing quality, simplifying the model and improving accuracy while reducing computational overhead. Various algorithms, including Support Vector Machines (SVM), Naïve Bayes, and Artificial Neural Networks (ANN), are evaluated using metrics like accuracy, precision, recall, and F1 score. This approach demonstrates the potential of data-driven methods to provide consistent, cost-effective quality assessments, enhancing operational efficiency and enabling more reliable quality control in the wine industry.

II. LITERATURE REVIEW

The application of machine learning to predict wine quality has garnered significant interest as a data-driven method to enhance quality control and optimize production processes in the wine industry. Traditional quality assessments rely on sensory evaluations by experts, which, while effective, are subjective and resource-intensive. In contrast, machine learning models offer an objective, reliable, and efficient alternative by analyzing chemical attributes of wine. Research in this area covers a range of topics, including feature selection, classification algorithms, and evaluation metrics, all contributing to the development of accurate and interpretable prediction models.

A pivotal study introduced a widely-used dataset for wine quality prediction, emphasizing attributes such as acidity, residual sugar, pH, chlorides, and alcohol content. This research demonstrated that machine learning algorithms like decision trees, neural networks, and support vector machines could reasonably approximate expert sensory ratings. Building on these findings, subsequent studies have focused on refining models, with feature selection emerging as a key factor in improving model performance. By identifying and prioritizing critical attributes, such as alcohol and

DOI: 10.48175/568

acidity, researchers have enhanced model efficiency and accuracy while simplifying computations. Additionally, advancements in machine learning techniques, including ensemble methods and deep learning, have further expanded the potential of predictive models. Metrics like accuracy, precision, recall, and F1 score remain central to evaluating these models, ensuring balanced performance across varying quality levels. Together, these innovations highlight the transformative role of machine learning in streamlining wine quality assessment, offering scalable, data-driven solutions for the industry.

III. METHODOLOGIES

Data Preprocessing:

Data preprocessing is a fundamental step in this project to prepare the wine quality dataset for effective machine learning modeling. The dataset, comprising various chemical attributes and quality ratings, undergoes systematic preprocessing to ensure data integrity and enhance model performance.

Handling Imbalanced Data

To address class imbalances in the quality ratings, the Synthetic Minority Over-sampling Technique (SMOTE) is utilized. SMOTE generates synthetic data points for underrepresented classes, balancing the class distribution. This approach ensures machine learning models do not disproportionately favor the majority class, enhancing predictive fairness.

Feature Scaling

Feature scaling is applied to standardize the dataset, ensuring consistent contributions from all features during model training. Initially, the **StandardScaler** normalizes features to a mean of 0 and a standard deviation of 1, optimizing the convergence of sensitive models. Subsequently, the **MinMaxScaler** scales features to a [0, 1] range, which benefits algorithms like Decision Trees and Random Forests that can be sensitive to feature magnitudes.

Feature Selection

To streamline the model and improve efficiency, feature selection is conducted using the **SelectKBest** method based on the ANOVA F-value. This technique evaluates the significance of each feature in relation to the target variable and selects the most impactful ones, refining the dataset for optimal performance.

Model Training:

Logistic Regression

A straightforward model, Logistic Regression is efficient for multi-class classification. It predicts the probability of each class using a linear decision boundary, offering simplicity and speed.

Decision Tree Classifier

This non-linear model constructs a tree-like structure to make decisions based on feature values. While interpretable, it requires careful pruning to avoid overfitting.

Random Forest Classifier

As an ensemble method, Random Forest combines multiple Decision Trees, averaging their outputs to enhance robustness and reduce overfitting. This approach often delivers high accuracy and adaptability.

Model Evaluation:

Model performance is assessed using accuracy metrics on test data, complemented by cross-validation. Cross-validation splits the training data into folds, training and validating iteratively to ensure the model generalizes well across diverse data subsets.

DOI: 10.48175/568

Results Interpretation

Accuracy metrics from test data and cross-validation provide insights into model efficacy. This analysis aids in selecting the most suitable model for practical applications.

Additional Approaches:

Support Vector Machines (SVM)

SVM creates a hyperplane to separate classes effectively, excelling in binary and multi-class classification. Techniques like One-vs-One and One-vs-All are employed for multi-class scenarios, leveraging feature scaling for improved accuracy.

Naïve Bayes

A simple yet powerful classifier based on Bayes' Theorem, Naïve Bayes assumes feature independence. It excels in handling large datasets and performs multiclass predictions with minimal computation time.

Multi-Layer Perceptron (MLP)

MLP is a neural network model with one or more hidden layers. Its ability to learn complex, non-linear relationships makes it well-suited for predicting wine quality. Careful tuning of hidden layers significantly impacts accuracy, demonstrating its sensitivity and potential.

Random Forest Algorithm

Random Forest employs bagging to construct diverse decision trees and aggregates their predictions. Its robustness against overfitting and ability to handle feature variability make it a reliable choice for wine quality prediction.

By applying these preprocessing techniques and models, this project seeks to accurately classify wine quality, leveraging the strengths of each approach to derive insightful and actionable outcomes.

IV. RESULT

The performance of the Logistic Regression, Decision Tree, and Random Forest classifiers was evaluated using the wine quality dataset. Each model's accuracy and cross-validation performance were analyzed to determine its effectiveness in predicting wine quality. The results are summarized below:

1. Logistic Regression

- **Accuracy**: 56.48%
- **Cross-Validation Scores**: [0.5566, 0.5780, 0.5841, 0.5559, 0.5804]
- **Mean Cross-Validation Accuracy**: 57.10%
- **Observations**: Logistic Regression exhibited the lowest accuracy among the models. While it performed as expected for a baseline linear model, its inability to capture non-linear relationships in the dataset limited its performance.

2. Decision Tree

- **Accuracy**: 78.24%
- **Cross-Validation Scores**: [0.8073, 0.7905, 0.7599, 0.7871, 0.7550]
- **Mean Cross-Validation Accuracy**: 77.99%
- **Observations**: The Decision Tree model significantly outperformed Logistic Regression, capturing non-linear relationships in the data. However, it exhibited a slight tendency to overfit, evident from its relatively higher standalone accuracy compared to the cross-validation mean.

3. Random Forest

- **Accuracy**: 86.92%
- **Cross-Validation Scores**: [0.8685, 0.8425, 0.8502, 0.8545, 0.8698]
- **Mean Cross-Validation Accuracy**: 85.71%
- **Observations**: Random Forest demonstrated the best performance among the three models, achieving the highest accuracy and mean cross-validation score. Its ensemble approach effectively reduced overfitting and enhanced generalization, making it the most reliable model for this task.

DOI: 10.48175/568

Table I Methodologies used

MODEL	ACCURACY	CROSS VALIDATION SCORES	MEAN CV ACCURACY
Logistic regression	0.5648	[0.5566, 0.5780, 0.5841, 0.5559, 0.5804]	0.5710
Decision tree	0.7824	[0.8073, 0.7905, 0.7599, 0.7871, 0.7550]	0.7799
Random forest	0.8692	0.8685, 0.8425, 0.8502, 0.8545, 0.8698]	0.8571

Feature Importance

Feature importance analysis using the Random Forest classifier highlighted **alcohol**, **volatile acidity**, and **sulphates** as the most significant predictors of wine quality. These features align with established domain knowledge, underscoring their critical role in determining wine quality.

The Random Forest classifier emerged as the most effective model for predicting wine quality, demonstrating superior accuracy and generalizability compared to Logistic Regression and Decision Tree models. The use of SMOTE to address class imbalance and feature selection to optimize the dataset contributed significantly to these results. Logistic Regression, while limited in its performance, served as a useful baseline, and Decision Tree provided insights into feature interactions despite slight overfitting. These findings highlight the potential of Random Forest for automating wine quality assessment and pave the way for future research into more advanced ensemble methods and hybrid models for enhanced predictive accuracy.

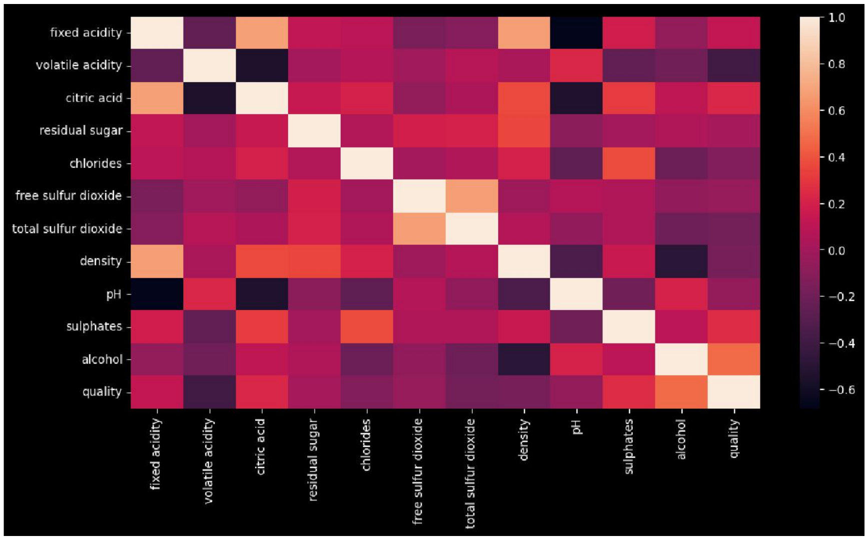

Fig 1 Correlation heat map

The correlation heat map for wine quality prediction offers insights into how different chemical properties relate to quality ratings. A moderate positive correlation often appears between alcohol content and quality, indicating that wines with higher alcohol levels tend to receive better quality ratings. In contrast, volatile acidity typically has a strong negative correlation with quality, suggesting that higher levels of volatile acidity, which can lead to unpleasant flavors, are associated with lower quality. Other factors, such as sulphates and pH, may show weaker correlations, with sulphates sometimes having a slight positive impact on quality due to their preservative properties. Attributes like residual sugar, fixed acidity, and chlorides generally exhibit low correlation with quality, implying they have less direct influence on ratings. This information can guide feature selection for predictive modeling, highlighting the importance of alcohol and volatile acidity as key predictors of wine quality.

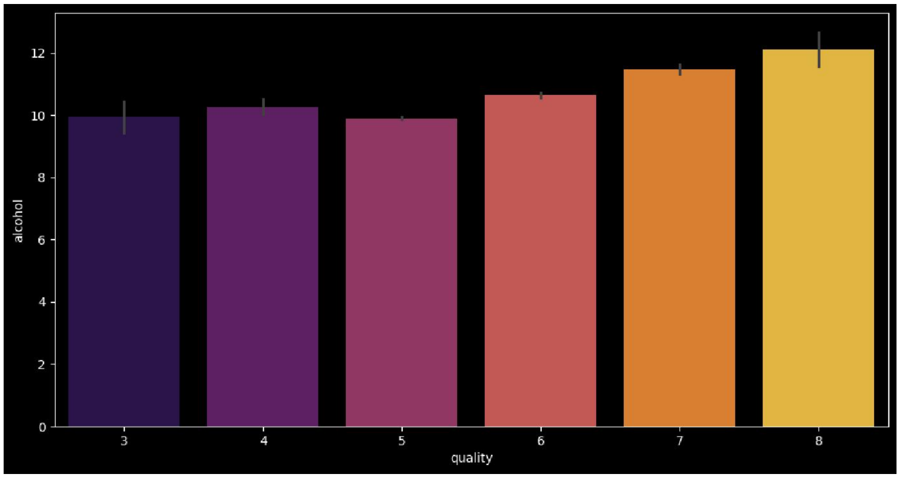

Fig 2. Bar plot

The bar plot of alcohol content by wine quality (output.png) reveals a trend where wines with higher quality ratings tend to have higher average alcohol content. As quality increases, there is a noticeable rise in the alcohol level, particularly for the highest-rated wines. This suggests that alcohol content might play a role in perceived wine quality, aligning with previous findings from the correlation analysis.

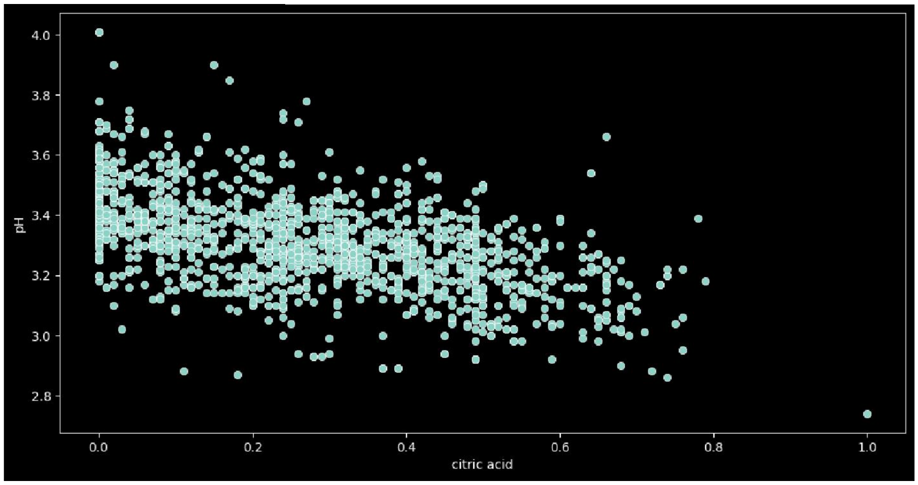

Fig 3 Scatter plot of pH vs Citric acid

The scatter plot of pH versus citric acid shows an inverse relationship, where higher levels of citric acid are generally associated with lower pH values. This pattern aligns with the chemical nature of citric acid as an acidifier, which lowers pH in wine. The trend suggests that as the citric acid content increases, the wine becomes more acidic, as reflected by a lower pH.

V. CONCLUSION

This project explored the use of basic machine learning models, including logistic regression and decision trees, to predict wine quality based on chemical attributes like acidity, alcohol content, pH, and sugar levels. The models found that alcohol content and volatile acidity were key factors, with alcohol being the strongest predictor. While the decision tree performed better by capturing non-linear relationships, the overall accuracy was limited due to the simplicity of the models and the absence of more complex interactions. The project highlighted the importance of certain chemical features in wine quality but also showed the need for more sophisticated models to improve prediction accuracy.

VI. FUTURE ENHANCEMENTS

To improve the wine quality prediction model, more advanced algorithms such as Random Forests, Gradient Boosting Machines (GBM), and Neural Networks could enhance predictive power by handling complex relationships and minimizing overfitting. Additional features like vineyard region, grape variety, and wine age could provide more context for predictions, while interaction features between chemical attributes could uncover new relationships. Hyperparameter tuning, cross-validation, and data preprocessing improvements, including stratified k-fold and nested cross-validation, would optimize model performance. Additionally, tools like SHAP and LIME could improve model interpretability, and expanding the dataset to include diverse wine samples would enhance generalization. Finally, deploying the model as a web or mobile application could provide practical value for the wine industry, enabling real-time quality predictions and decision-making.

REFERENCES

[1]. https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/

[2]. https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction

[3]. https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/

[4]. https://www.nature.com/articles/s41598-023-44111-9