IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

Machine Learning-Based Predictive Modelling for the Enhancement of Wine Quality

Vishal Deshmukh[1] and Riya Tandel[2]
Assistant Professor Department of IT[1]
Student, P.G. Department of IT[2]
Veer Wajekar ASC College, Phunde, Uran

Abstract: *The certification of wine quality is crucial for the wine industry to ensure product standardization, consumer safety, and market competitiveness. This research proposes the application of multiple machine learning (ML) models to predict wine quality using the Red Wine Dataset (RWD), which contains 11 physicochemical attributes. Models including Random Forest (RF), Decision Trees (DT), AdaBoost, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) are evaluated for their predictive performance. Notably, XGBoost and RF demonstrated the highest accuracy. Feature selection and cluster analysis were performed to identify key attributes and manage collinearity. The study supports ML-driven quality certification and provides insights into essential physicochemical parameters of wine quality.*

Keywords: Wine quality prediction, Machine Learning, Feature selection, Random Forest, XGBoost, Data preprocessing, Cluster analysis

I. INTRODUCTION

Wine is a globally consumed beverage whose quality has significant implications on its commercial value and consumer satisfaction. In recent years, machine learning has emerged as a robust method to automate and enhance the wine quality certification process by analyzing physicochemical characteristics. This study utilizes the Red Wine Dataset (RWD), which comprises 1599 samples and 11 key features, to build ML models capable of accurately predicting wine quality. The ultimate goal is to minimize subjective errors from sensory evaluations and develop objective, reproducible quality assessments.

Related Work Past studies have explored ML techniques in wine quality prediction. Cortez et al. (2009) employed regression models on white wine data, achieving high predictive accuracy. Aich et al. (2020) used PCA and decision tree classifiers to select critical features. Gupta et al. (2021) combined Random Forest with KNN to rate wine as Good, Average, or Bad. Kumar et al. (2019) compared RF, SVM, and Naive Bayes, while Shaw et al. (2020) concluded that RF yielded the best results. These studies highlight the relevance and potential of ML in the wine industry but also expose gaps such as limited datasets and lack of generalization. Materials and Methods 3.1 Dataset Description The RWD from the UCI Machine Learning Repository contains 1599 samples with the following features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and wine quality (rated from 3 to 8). For this study, we transformed the multiclass labels into binary classification (good vs. bad quality). Data Preprocessing Initial preprocessing involved handling missing values and duplicate records. The data was normalized using standard scaling (mean = 0, standard deviation = 1). Class balancing was analyzed; since the dataset was not significantly imbalanced, SMOTE was not applied. Exploratory Data Analysis (EDA) was conducted using heatmaps and correlation matrices. Feature Selection Feature importance was assessed using RF and XGBoost. The top three features (alcohol, volatile acidity, and sulphates) were consistently ranked highest. A stepwise reduction method was employed to test the impact of feature selection on model accuracy.

Cluster Analysis Hierarchical clustering and k-means clustering were applied to group similar attributes and minimize multicollinearity. Features with high correlation were clustered and redundant predictors were removed, improving model performance and reducing computational load.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

Machine Learning Models Five models were implemented:

- **Decision Tree (DT):** Simple and interpretable but prone to overfitting.
- **Random Forest (RF):** Robust against overfitting; uses multiple decision trees.
- **AdaBoost:** Combines weak learners by emphasizing misclassified data points.
- **Gradient Boosting (GB):** Sequentially adds models to minimize loss function.
- **Extreme Gradient Boosting (XGBoost):** Advanced GB implementation with regularization and parallel processing.

Hyperparameter tuning was conducted using GridSearchCV for RF and XGBoost to optimize model parameters.
Results and Discussion 5.1 Model Accuracy

- XGBoost (with selected features): 100%
- Random Forest (with all features): 98.4%
- RF (with selected features): 99.1%
- AdaBoost: 93.2%
- Gradient Boost: 94.7%
- DT: 89.3%

Feature Importance XGBoost and RF highlighted the following key features:
Alcohol Volatile Acidity Sulphates These features have strong biochemical relevance, aligning with oenological standards.
Cluster Analysis Outcome Redundant variables such as total sulfur dioxide and density were grouped and pruned. Clustering improved model interpretability and reduced overfitting risk.
Comparative Analysis This study surpasses prior work by using a larger dataset, robust preprocessing, and model tuning. Unlike earlier works, it emphasizes cluster analysis and interpretable feature selection.

II. CONCLUSION

Machine learning models, particularly XGBoost and Random Forest, can accurately predict wine quality using physicochemical data. Feature selection and clustering further enhance performance and interpretability. These findings advocate for integrating ML-based systems into wine certification processes.
Future Work Future studies can expand the model to white wines, include sensory attributes, and explore deep learning methods like CNNs and LSTMs. Real-time quality assessment systems can be developed using edge-AI.

REFERENCES

[1]. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553.
[2]. Aich, S., Park, J. H., & Kim, H. C. (2020). Machine Learning-based Wine Quality Prediction: Comparative Study. Sensors, 20(21), 6055.
[3]. Gupta, S. K., et al. (2021). Machine Learning-based Wine Classification System. Procedia Computer Science, 185, 384–391.
[4]. Shaw, R., & Patra, S. (2020). Performance analysis of classification models for wine quality prediction. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(2).
[5]. Kumar, S., et al. (2019). Comparative Study of Wine Quality Prediction using ML Techniques. Journal of Data Mining and Knowledge Management Process, 9(1), 1–9.
[6]. Bhardwaj, A., et al. (2022). Predicting Wine Quality Using Chemical Properties: A Machine Learning Approach. Journal of Food Quality, 2022.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

[7]. Tiwari, A., et al. (2023). A Conceptual Framework for Sensory and Chemical Data Analysis in Wine Quality Prediction. Applied Sciences, 13(4), 2051.

[8]. Ma, J., et al. (2020). Wine Type Classification Based on Physicochemical Data Using Deep Learning. Computers and Electronics in Agriculture, 169.

[9]. Prez, L., et al. (2021). PCA and SVM-Based Classification of Wines. Journal of Food Engineering, 109(4).

[10]. Mahima, S., & Kumar, V. (2021). Quality Assessment of Wine Using Ensemble Machine Learning Methods. International Journal of Advanced Computer Science and Applications, 12(3), 64–71.