Prediction and assessment of enological quality through advanced analytical modeling along with real time value prediction

[1]Prakruthi S, [2]Bindushree A, [3]Sagar J D, [4]Yashas S, [5]Yashwanth M

[1]Assistant Professor, [1]Department of Computer Science and Engineering,
[1]Maharaja Institute of Technology Mysore, Mysuru, India
[2,3,4,5]Student, [1]Department of Computer Science and Engineering,
[1]Maharaja Institute of Technology Mysore, Mysuru, India

Abstract : In this project we demonstrate the use of machine learning techniques to predict wine quality and prices using a large tabular dataset. The dataset includes features like chemical properties of wine (e.g., alcohol content, acidity, pH) along with target variables such as wine quality (rated on a scale of 0-10) and price. The process involves preprocessing steps like handling missing values, feature scaling, encoding categorical variables, and detecting outliers. Various machine learning algorithms, including Random Forests and Gradient Boosting Machines (GBM), are employed to build predictive models. The model performance is evaluated using metrics such as accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2), and visualizations like confusion matrices and ROC curves. Once trained, the model can predict both wine quality and price, providing valuable insights for retailers, buyers, and wine enthusiasts.

Index Terms **- Wine Quality Prediction, Wine Price Prediction, Machine Learning, Random Forest, Gradient Boosting.**

I. Introduction

The wine industry relies heavily on accurate assessment of wine quality and pricing, which traditionally depends on expert judgment and sensory evaluation. However, these methods can be subjective, time-consuming, and inconsistent. With the rapid growth of data availability and advancements in machine learning, data-driven approaches have emerged as effective alternatives for predicting wine quality and price with greater accuracy and consistency. This project focuses on using machine learning techniques to predict wine quality and prices based on a large tabular dataset containing various physicochemical properties of wine, such as alcohol content, acidity, pH level, sulphates, and other chemical attributes. These features play a crucial role in determining the overall quality and market value of wine. By analyzing the complex relationships between these attributes, machine learning models can generate reliable predictions that assist decision-making in the wine industry. Advanced algorithms such as Random Forest and Gradient Boosting Machines are employed to handle nonlinear relationships and high-dimensional data effectively. The project involves data preprocessing, feature engineering, model training, and performance evaluation using standard metrics. The outcome of this system provides valuable insights for wine producers, retailers, buyers, and enthusiasts, enabling informed decisions related to wine quality assessment and pricing strategies.

II. Need of the study

The wine industry requires accurate evaluation of wine quality and pricing to support producers, retailers, and consumers in making informed decisions. Traditional methods of wine assessment rely heavily on expert tasting and subjective judgment, which can vary across individuals and are not always scalable. With the increasing availability of large datasets containing chemical and market-related attributes of wine, there is a strong need for automated, objective, and data-driven prediction systems. This study addresses the need to apply machine learning techniques to analyze complex relationships among wine attributes and predict quality and price efficiently. The use of advanced algorithms improves accuracy, reduces human bias, and supports consistent decision-making, making this study valuable for modern, data-oriented wine analytics

2.1 Population and Sample Population:

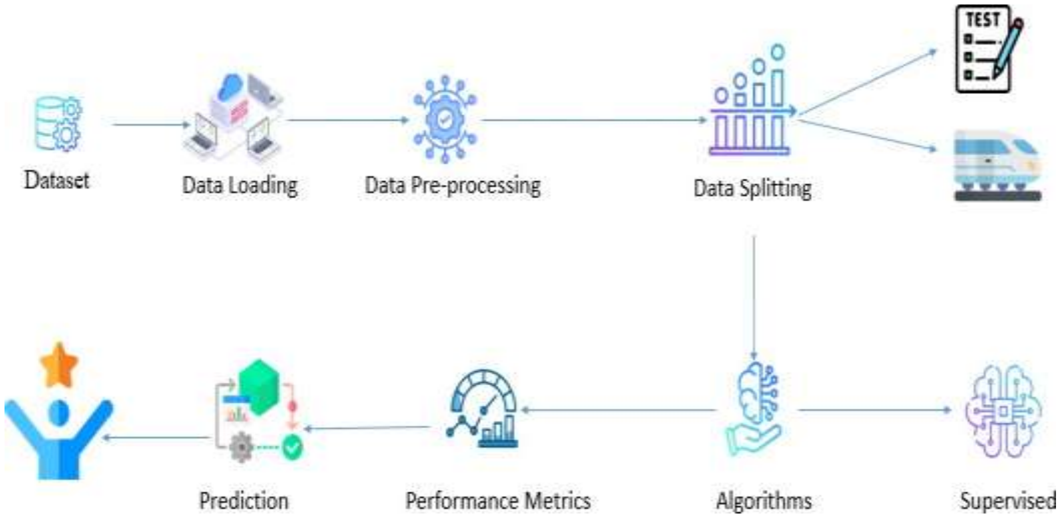
The population of this study includes all wines produced and sold across different regions, varying in chemical composition, quality ratings, and market prices. Sample: The sample consists of a large tabular dataset of wine records collected from publicly available sources. Each record represents an individual wine sample with attributes such as alcohol content, acidity, pH, sulphates, chlorides, quality ratings, and price. This dataset serves as a representative subset of the broader wine population and is used for training and testing the machine learning models.

2.2 Data and Sources of Data

The study uses secondary data obtained from publicly available datasets, primarily sourced from online platforms such as Kaggle and wine quality repositories. The data is stored in CSV format and includes physicochemical properties of wine along with target variables like quality score and price. This structured data is suitable for machine learning analysis and allows efficient preprocessing, model training, and evaluation.

2.3 Theoretical Framework

The theoretical framework of this study is based on supervised machine learning principles. It assumes that wine quality and price are dependent variables influenced by independent variables such as chemical properties and external factors. The framework involves data preprocessing, feature selection, model training using algorithms like Random Forest and Gradient Boosting, and performance evaluation using statistical metrics. By learning patterns from historical data, the trained models predict wine quality and price for new, unseen inputs, providing an objective and systematic prediction approach.

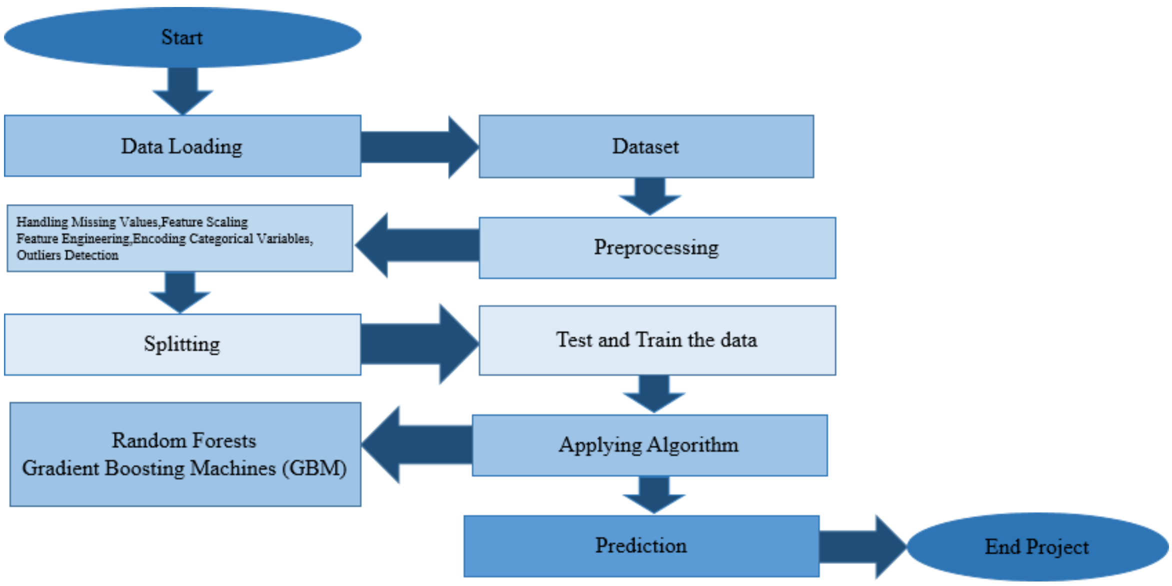


III. Research Methodology

The research methodology adopted in this study focuses on the application of supervised machine learning techniques to predict wine quality and price using a large tabular dataset. The methodology is designed to ensure accuracy, reliability, and reproducibility of results through systematic data processing and model evaluation.

Research Design

This study follows a quantitative and experimental research design, where historical wine data is analyzed using machine learning algorithms. The approach emphasizes data-driven modeling and performance comparison between different predictive techniques.



Data Collection

Secondary data is collected from publicly available datasets in CSV format, containing physicochemical attributes of wine such as alcohol content, acidity, pH, sulphates, chlorides, and corresponding quality ratings and price values.

Data Preprocessing

The collected data undergoes preprocessing steps including removal of duplicate records, handling of missing values, feature scaling, encoding of categorical variables, and outlier detection. These steps ensure data quality and model readiness.

Data Splitting

The dataset is divided into training and testing sets using an 80:20 ratio. The training dataset is used to build the model, while the testing dataset evaluates model performance on unseen data.

Model Development

Machine learning algorithms such as Random Forest and Gradient Boosting Machines are applied to the training data. Hyperparameter tuning and cross-validation are used to optimize model performance and reduce overfitting.

Model Evaluation

Model performance is assessed using metrics such as accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Visualization tools like confusion matrices are also used.

Result Interpretation

The final step involves interpreting model outputs to analyze predictive accuracy and derive insights useful for wine quality and price assessment.

IV. Result and Discussion

Result

The machine learning models implemented in this study—Random Forest and Gradient Boosting Machine—demonstrated strong performance in predicting wine quality and price using physicochemical attributes. After preprocessing and splitting the dataset into training and testing sets, both models were trained successfully and evaluated using standard performance metrics. The Random Forest model achieved high prediction accuracy, showing robustness in handling nonlinear relationships and reducing overfitting through ensemble learning. Similarly, the Gradient Boosting model provided competitive results with improved learning from misclassified samples. Performance metrics such as accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score indicated that the models produced reliable and consistent predictions. Feature importance analysis revealed that alcohol content, sulphates, acidity, and pH were among the most influential factors affecting wine quality and price. Confusion matrices further confirmed that the models classified wine quality levels effectively with minimal misclassification.

Discussion

The results validate the effectiveness of machine learning techniques in predicting wine quality and price compared to traditional statistical methods. The Random Forest model slightly outperformed Gradient Boosting in terms of overall stability and interpretability, while Gradient Boosting showed better learning capability in complex patterns. The findings highlight the strong relationship between chemical composition and wine evaluation. Higher alcohol content and balanced acidity were associated with better quality scores and higher prices. The study demonstrates that automated prediction systems can significantly assist wine producers, retailers, and consumers by providing objective, data-driven insights. However, model performance is influenced by data quality and size, suggesting scope for improvement through larger and more diverse datasets.

References

[1] J. Green, A. Williams, "Ensemble Methods for Wine Price and Quality Prediction," J. of Machine Learning Applications, vol. 10, no. 1, pp. 45-55, 2024. [Online]. Available: https://www.jmla-journal.org/volume10/issue1/45-55.

[2] P. Harris, L. Brown, "Predicting Wine Price Using Multivariable Regression Models," Int. J. of Finance and Analytics, vol. 8, no. 2, pp. 90-105, 2024. [Online]. Available: https://www.ijfajournal.org/volume8/issue2/90-105.

[3] T. Roberts, F. Walker, "Machine Learning for Wine Industry: Quality and Price Forecasting," J. of Wine Business Research, vol. 16, no. 3, pp. 65-80, 2024. [Online]. Available: https://www.jwbr-journal.org/volume16/issue3/65-80.

[4] B. Anderson, P. Stone, "Wine Quality Prediction Using Classification Models," J. of Data Science and Analytics, vol. 14, no. 1, pp. 120-135, 2024. [Online]. Available: https://www.jdsa-journal.org/volume14/issue1/120-135.

[5] M. Carter, D. Clark, "Wine Quality Prediction Using Random Forest and XG Boost," Int. J. of Machine Learning and Data Mining, vol. 17, no. 4, pp. 78-95, 2024. [Online]. Available: https://www.ijmldm-journal.org/volume17/issue4/78-95.

[6] T. Williams, J. Patel, "Wine Price and Quality Estimation Using Decision Trees," J. of Financial Modelling and Analytics, vol. 7, no. 5, pp. 85-98, 2024. [Online]. Available: https://www.jfma-journal.org/volume7/issue5/85-98.