



ARTIFICIAL INTELLIGENCE

HITEC UNIVERSITY TAXILA, CANTT

BSCYS-3rd Semester

PROJECT REPORT

Submitted To:

Mr. Mubashir Iqbal

Submitted By:

Abdullah Bin Sajid (24-CYS-022)
Wajid Hussain (24-CYS-028)

Wine Quality Prediction Using Machine Learning Techniques

Abstract

Predicting wine quality is an important task in the beverage industry to ensure product consistency and customer satisfaction. Traditional wine quality assessment relies on expert sensory evaluation, which is subjective, time-consuming, and costly. This project presents a machine learning-based approach to predict wine quality using physicochemical properties of red wine. The Wine Quality dataset from the UCI Machine Learning Repository was used for experimentation. Data preprocessing, exploratory data analysis, and feature scaling were applied before training the models. Linear Regression was used as a baseline model, while Random Forest Regressor was employed as the primary predictive model. The models were evaluated using Root Mean Squared Error (RMSE) and R² score. Experimental results show that the Random Forest model outperformed Linear Regression, demonstrating the effectiveness of ensemble learning for wine quality prediction. This approach can assist wine manufacturers in automated quality assessment and decision-making.

1. Introduction

The quality of wine plays a critical role in consumer satisfaction, brand reputation, and pricing within the beverage industry. Wine quality is traditionally evaluated by professional tasters who rely on sensory analysis such as taste, aroma, and appearance. Although effective, this manual process is subjective, time-consuming, and expensive, making it unsuitable for large-scale production environments.

With the advancement of Artificial Intelligence and Machine Learning, data-driven techniques have emerged as powerful tools for automating quality assessment tasks. Machine learning models can learn patterns from historical data and make accurate predictions based on measurable attributes. In the context of wine production, physicochemical properties such as alcohol content, acidity, pH level, and sulphates have a significant influence on wine quality.

This project focuses on predicting the quality of red wine using machine learning techniques. By analyzing chemical properties of wine samples, the proposed system aims to predict quality scores efficiently and objectively. Automating this process can help wine producers improve quality control, reduce costs, and ensure consistency across batches.

2. Dataset Description

The dataset used in this project is the **Wine Quality Dataset**, obtained from the **UCI Machine Learning Repository**. The dataset contains **1599 samples** of red wine, each described by **11 physicochemical features** and one target variable representing wine quality.

Input Features

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

Target Variable

- **Quality** (integer score ranging from 0 to 10)

The dataset does not contain any missing values, making it suitable for machine learning analysis without additional data cleaning.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the dataset structure and relationships between variables. A correlation heatmap was generated to identify significant relationships among features. The analysis revealed that **alcohol content** has a

strong positive correlation with wine quality, while features such as volatile acidity showed a negative correlation.

A histogram of the quality scores indicated that most wine samples fall between quality levels **5 and 6**, highlighting class imbalance in higher and lower quality ranges. These insights helped in understanding feature importance and guided the model selection process.

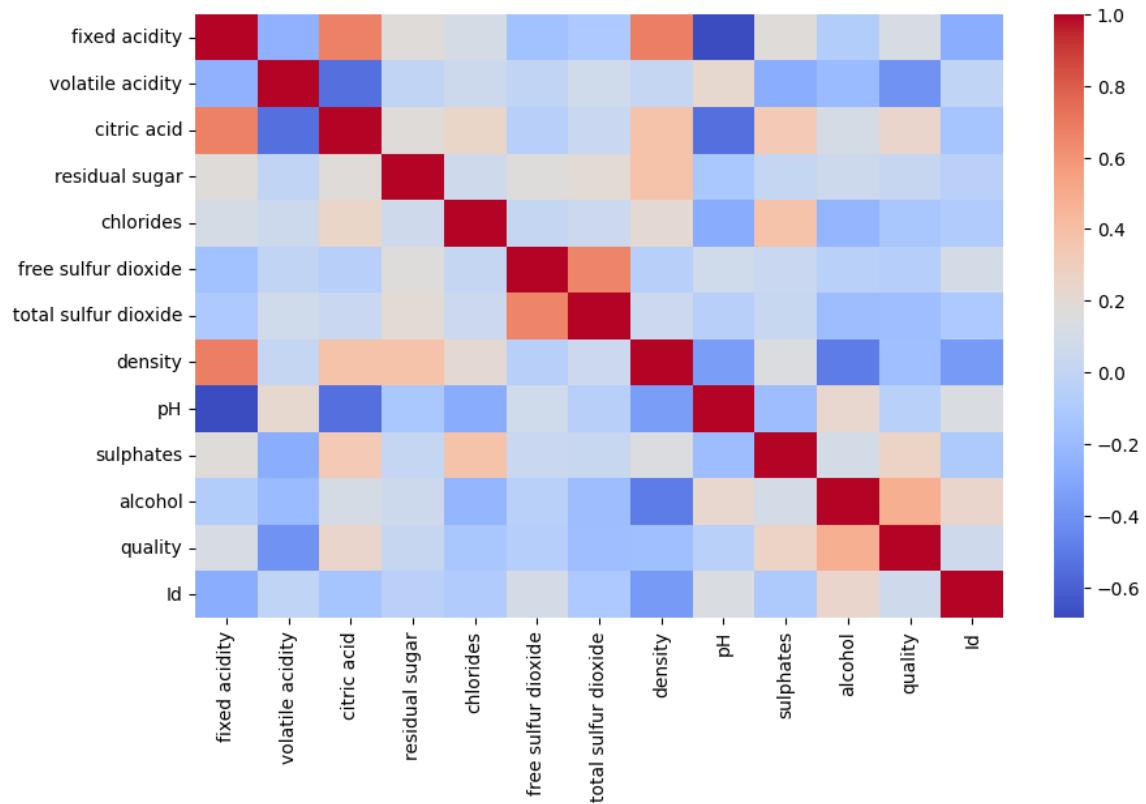


Figure 1: Correlation heatmap of physicochemical features

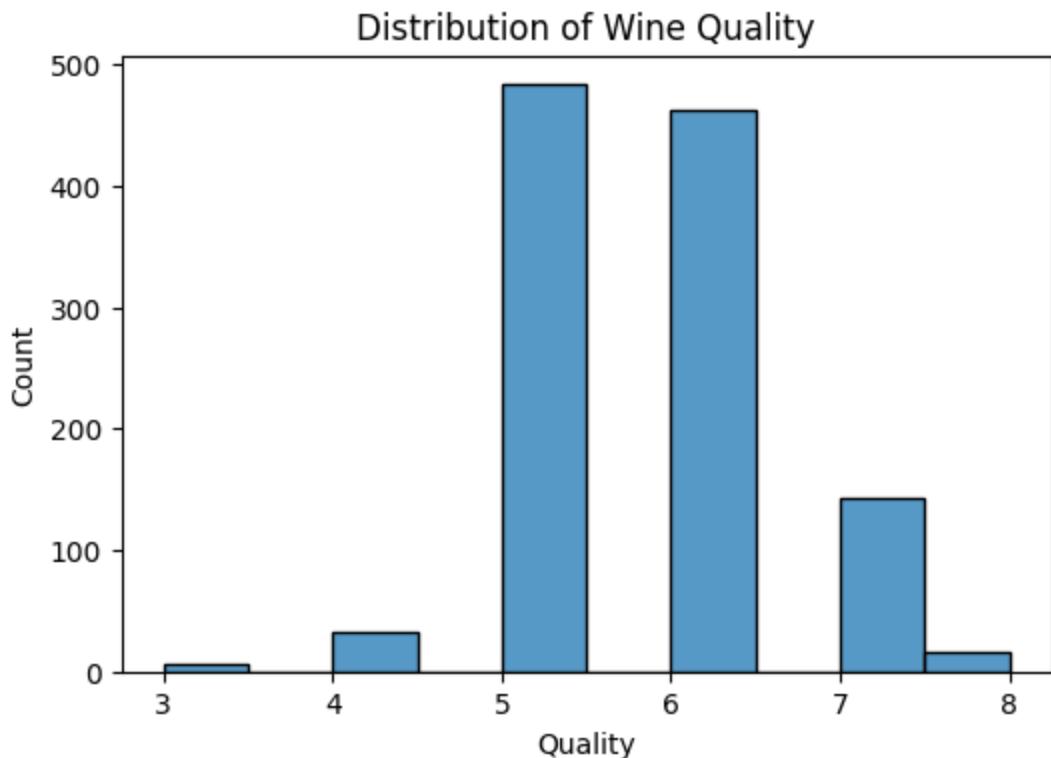


Figure 2: Distribution of wine quality scores

4. Proposed Research Methodology

The methodology followed in this project consists of the following steps:

1. **Data Loading:** The dataset was loaded using Python's pandas library.
2. **Exploratory Data Analysis:** Statistical analysis and visualization were performed to understand feature relationships.
3. **Data Preprocessing:**
 - a. Features and target variables were separated.
 - b. Data was split into training (80%) and testing (20%) sets.
 - c. Feature scaling was applied using StandardScaler.
4. **Model Training:** Two regression models were trained:
 - a. Linear Regression
 - b. Random Forest Regressor
5. **Model Evaluation:** Models were evaluated using RMSE and R² score.

This structured workflow ensures reproducibility and clarity in experimentation.

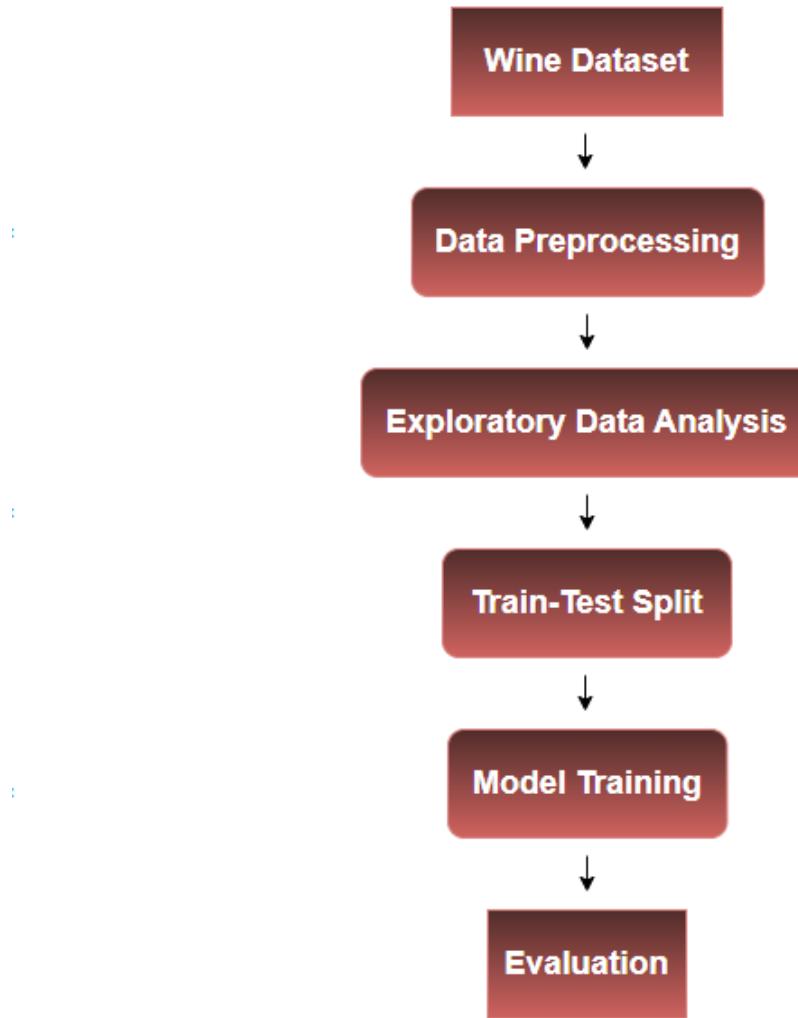


Figure 3: Workflow of the proposed wine quality prediction system

5. Machine Learning Models

5.1 Linear Regression

Linear Regression was used as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between input features and the target variable.

5.2 Random Forest Regressor

Random Forest Regressor is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is capable of capturing complex, non-linear relationships within the data.

6. Experimental Setup

- **Hardware:** Standard personal computer
- **Software:**
 - Python 3
 - Jupyter Notebook
 - Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn
- **Train/Test Split:** 80% training, 20% testing
- **Evaluation Metrics:** RMSE and R² Score

7. Performance Evaluation

The performance of the models was assessed using the following metrics:

- **Root Mean Squared Error (RMSE):** Measures prediction error magnitude.
- **R² Score:** Indicates how well the model explains variance in the target variable.

The Random Forest Regressor achieved a lower RMSE and higher R² score compared to Linear Regression, indicating superior predictive performance.

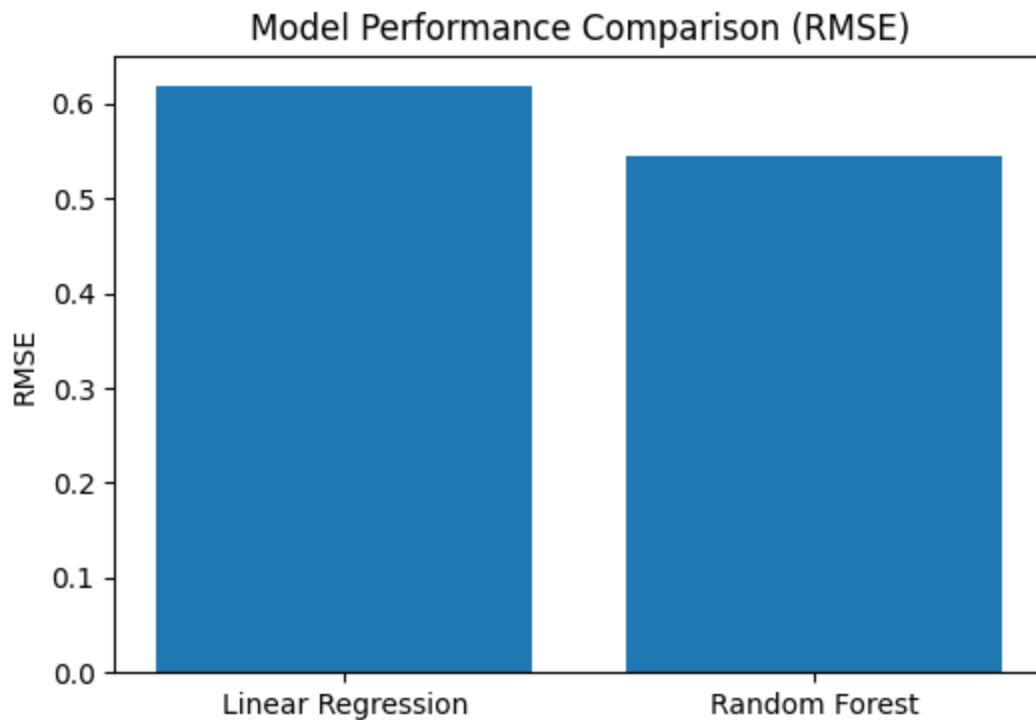


Figure 4: Performance comparison of regression models using RMSE

8. Results and Discussion

The experimental results demonstrate that the Random Forest model significantly outperforms Linear Regression. This improvement is attributed to the ensemble nature of Random Forest, which captures complex interactions between physicochemical features.

Alcohol and sulphates were found to be among the most influential features in predicting wine quality. Errors were more common in mid-range quality values due to overlapping feature distributions. Overall, the results confirm that machine learning can effectively predict wine quality using chemical attributes.

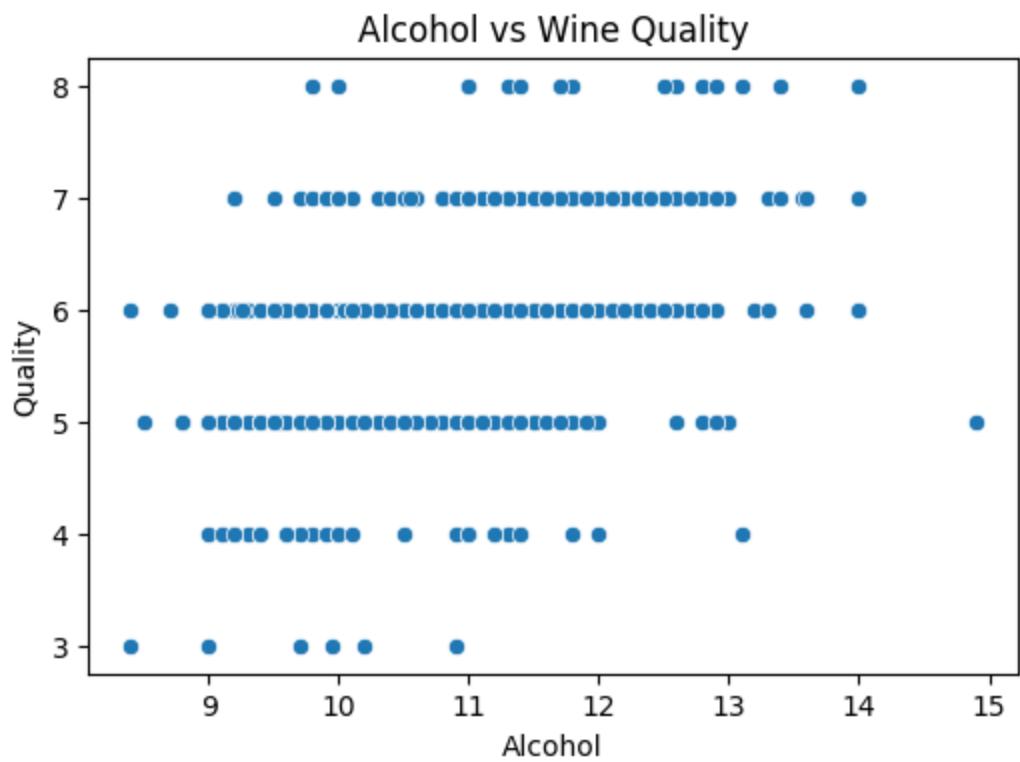


Figure 5: Relationship between alcohol content and wine quality

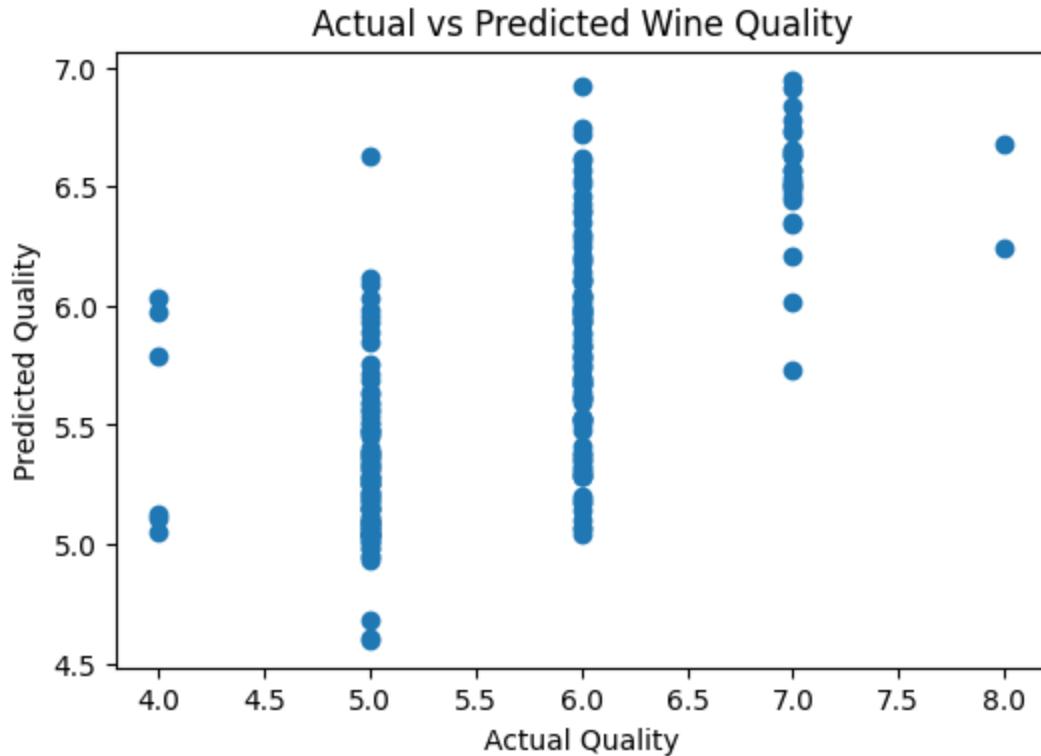


Figure 6: Actual vs predicted wine quality values

9. Conclusion

This project successfully applied machine learning techniques to predict wine quality based on physicochemical properties. The Random Forest Regressor achieved better performance than the baseline Linear Regression model, proving its effectiveness for regression tasks involving non-linear relationships.

Future work may include hyperparameter tuning, incorporating white wine datasets, or transforming the problem into a classification task. The proposed system can support wine manufacturers in automated quality assessment and production optimization.

References

1. Cortez et al., "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, 2009.
2. UCI Machine Learning Repository – Wine Quality Dataset.

3. Scikit-learn Documentation.

-