Final Report – Wine Quality Prediction using Linear Regression

By Varshini

1. Introduction

Wine quality assessment is an important task within the food and beverage domain, as it helps producers evaluate product consistency and market value. Traditionally, wine quality is judged by human sensory evaluation, which is subjective, time-consuming, and can vary significantly between evaluators.

The goal of this project is to **predict wine quality (a numerical score from 0 to 10)** using measurable physicochemical features in the well-known **Red Wine Quality dataset**. This project was developed as part of a machine learning hackathon emphasizing simple models, interpretability, and clear data-driven reporting.

The primary objective is to:

- Load and explore the dataset
- Train a **baseline Linear Regression model**
- Evaluate performance using RMSE and R^2
- Identify the most important predictive features
- Provide an analytical, clear explanation of outcomes

2. Dataset Overview

The dataset contains **1,599 samples**, each representing a red wine sample along with 11 measurable chemical attributes:

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide

- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

The target variable **quality** is a score (0–10), typically between **3 and 8**.

2.1 Data Quality Check

- No missing values were found, simplifying preprocessing.
- All features except quality are continuous numeric variables.
- No obvious data entry inconsistencies were detected.

2.2 Distribution Observations

- Majority of wines fall within quality scores **5, 6, and 7**.
- A slight imbalance exists but is acceptable for regression.
- Alcohol content and volatile acidity appear to have the strongest initial correlations with wine quality.

3. Exploratory Data Analysis (EDA)

3.1 Correlation Insights

A correlation heatmap revealed the following trends:

- **Alcohol** shows the strongest **positive** correlation with quality.
- **Volatile acidity** shows a clear **negative** correlation.
- **Sulphates** show a moderate positive correlation.
- **Density** is negatively correlated with alcohol and quality.
- Many features show only mild linear relationships, which suggests that simple linear regression may not fully capture the underlying complexity.

3.2 Feature Distributions

- Residual sugar, chlorides, and sulphates show right-skewed distributions.
- Density is tightly distributed around ~0.996–0.999 g/mL.
- Alcohol levels vary significantly, contributing to quality variation.

EDA suggests that wine quality is influenced by multiple weak-to-moderate chemical interactions, which aligns with expectations from real-world wine chemistry.

4. Preprocessing & Modeling Approach

4.1 Train-Test Split

The dataset was split into:

- **80% training set**
- **20% test set**

4.2 Feature Scaling

Linear Regression is sensitive to feature scale. Therefore, **StandardScaler** was used to standardize all predictor variables.

4.3 Model Choice

A **Linear Regression** model was chosen because:

- It is simple, interpretable, and quick to train
- Coefficients help identify the relative importance of features
- It aligns with the hackathon requirement for a baseline model
- It provides a solid foundation for comparing future models

5. Model Performance

5.1 Evaluation Metrics

Two metrics were used:

- **RMSE (Root Mean Squared Error)**
  Indicates how far predictions deviate from actual values on average.

- **R^2 (Coefficient of Determination)**
  Indicates what percentage of variance in wine quality is explained by the model.

5.2 Results

- **Test RMSE: 0.6870**
- **Test R^2: 0.3396**

5.3 Interpretation

- An RMSE of **0.6870** means predictions are, on average, within **±0.69** quality points of the true score.
- An R^2 of **0.3396** indicates the model explains roughly **34%** of the variation in wine quality.

Given that the dataset contains subjective human ratings and chemical interactions that may not be strictly linear, this performance is expected for a baseline model.

Models such as Random Forest, Gradient Boosting, or Lasso Regression would likely improve these metrics.

6. Most Influential Features

Using the absolute value of regression coefficients on scaled features, the top contributors to wine quality prediction were:

1. Residual sugar — positive effect

Wines with slightly higher residual sugar tended to receive higher predicted quality scores.
This may relate to the perception of balance and body in the wine.

2. Density — negative effect

Higher density wines generally scored lower in predicted quality.
Density often increases with higher sugar or lower alcohol content, which may contribute to lower sensory perception of quality.

3. Alcohol — positive effect

Alcohol content had a clear positive effect.
Higher alcohol levels often enhance mouthfeel and aromatics, aligning with real-world winemaking expectations.

These results align with known wine chemistry principles and provide confidence in model interpretability.

7. Conclusion

This project successfully developed a baseline predictive model for red wine quality using Linear Regression. Key achievements include:

- Thorough EDA to understand feature relationships
- Clean preprocessing with data scaling
- Clear model evaluation using RMSE and R^2
- Identification of top features influencing wine rating
- Preparation of a complete analytical report for submission

Although the model explains about one-third of the variance, this is a reasonable baseline given the complexity and subjectivity of wine quality assessment.

Future Improvements

To improve prediction accuracy, the following approaches are recommended:

- **Ridge and Lasso Regression** to handle multicollinearity
- **Random Forest or XGBoost Regression** for non-linear interactions
- **Hyperparameter tuning** for performance optimization
- **Cross-validation** to improve generalization
- **Feature engineering** (e.g., interaction terms, transformations)