DS 325 Final Project Report

Predicting Red Wine Quality from Chemical Properties

By: Pronob Sarker

Introduction

Wine quality assessment has traditionally been the domain of trained sommeliers who rely on subjective sensory evaluation—a process that, while valuable, lacks the objectivity and reproducibility of scientific measurement. The wine industry, with global revenue exceeding $340 billion annually, has significant economic incentives to understand the factors that objectively contribute to wine quality. Despite technological advances in winemaking, the relationship between measurable chemical properties and perceived quality remains incompletely understood. I was motivated to explore this relationship after observing the contrast between the scientific precision of modern wine production and the seemingly subjective nature of quality assessment. Could the chemistry in our glass truly predict the experience on our palate? I hypothesized that physicochemical properties of red wine (such as acidity, alcohol content, and sulphates) would significantly predict expert quality ratings, with certain properties having disproportionate influence. Using a dataset of 1,599 Portuguese red wine samples, I employed both linear regression and Random Forest models to quantify these relationships. My analysis revealed that approximately 52% of quality variation could be explained by chemical measurements alone, with alcohol content, volatile acidity, and sulphates emerging as the most influential predictors—findings that bridge the gap between wine science and sensory perception.

Methods

Dataset Description

This analysis utilized the Wine Quality dataset from the UCI Machine Learning Repository (Cortez et al., 2009), containing 1,599 red wine samples from Portuguese "Vinho Verde" wines. Each sample includes measurements of 11 physicochemical properties and a quality rating (0-10 scale) assigned by wine experts:

- **Fixed acidity**: Non-volatile acids (primarily tartaric acid)
- **Volatile acidity**: Primarily acetic acid (vinegar-like)

- **Citric acid**: Contributes freshness and flavor
- **Residual sugar**: Unfermented sugar content
- **Chlorides**: Salt content
- **Free/Total sulfur dioxide**: Preservatives
- **Density**: Related to sugar and alcohol content
- **pH**: Acidity level (typically 3-4)
- **Sulphates**: Additives contributing to sulfur dioxide levels
- **Alcohol**: Percent alcohol content
- **Quality**: Target variable (expert ratings)

Data Preprocessing

Initial exploration revealed a remarkably clean dataset with no missing values and only a small number of duplicates (approximately 1% of observations), which were removed. Several variables showed notable outliers, particularly in residual sugar and chlorides. Rather than removing these outliers, I retained them as they likely represent legitimate, if unusual, wine samples.

All features were standardized to have mean = 0 and standard deviation = 1 to facilitate comparison of feature importance and improve model performance. The distribution of quality ratings was approximately normal, concentrated in the 5-6 range, with no wines receiving the highest possible ratings (9-10).

Modeling Approach

I implemented two complementary modeling strategies:

1. **Linear Regression**: To establish baseline relationships and provide easily interpretable coefficients that directly quantify each property's influence on quality.

2. **Random Forest**: To capture potential non-linear relationships between chemical properties and wine quality. This ensemble method combines multiple decision trees trained on different subsets of data, improving prediction accuracy and providing robust feature importance measures.

Model validation employed 5-fold cross-validation to ensure reliable performance estimates. For the Random Forest model, I conducted hyperparameter tuning using GridSearchCV to optimize:

- Number of trees (n_estimators)
- Maximum tree depth (max_depth)

- Minimum samples required for splitting (min_samples_split)
- Minimum samples per leaf (min_samples_leaf)

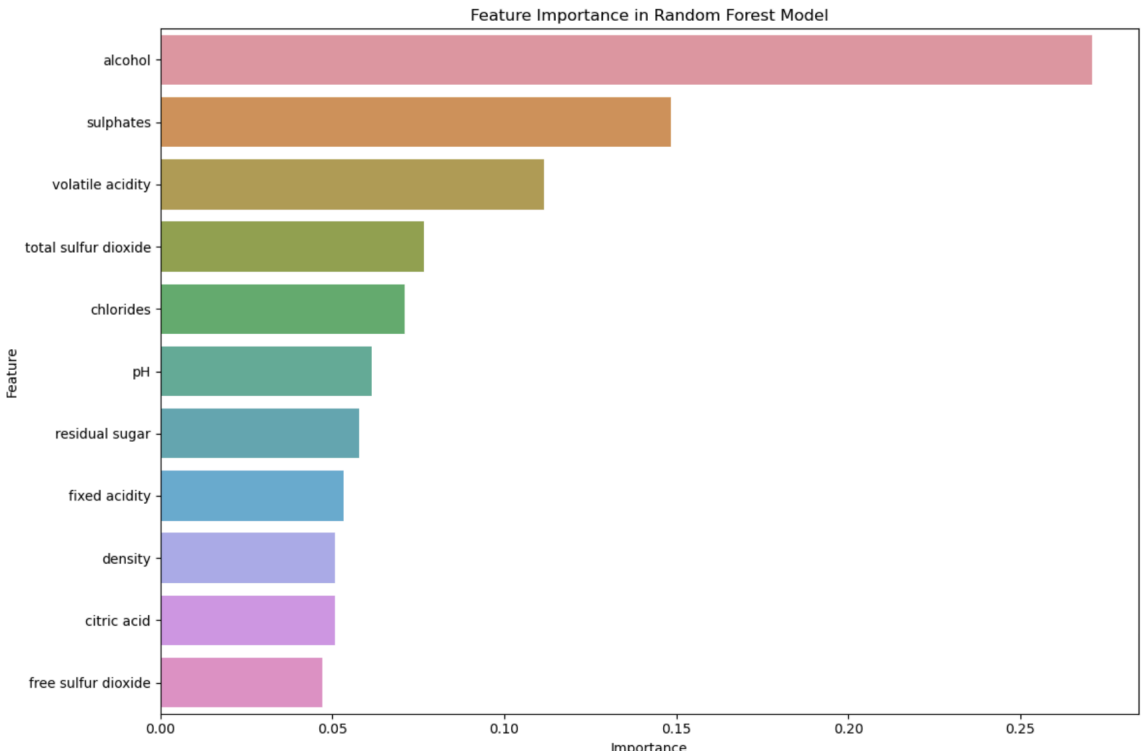
Model performance was evaluated using coefficient of determination (R^2), root mean square error (RMSE).

Results

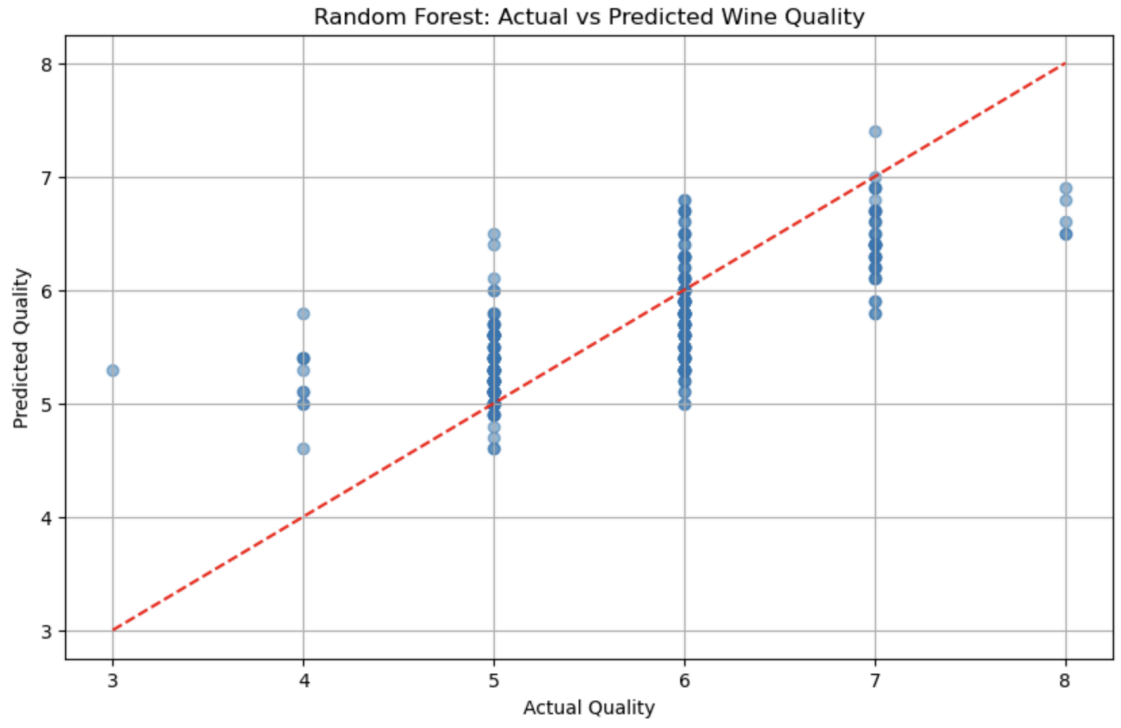
The Random Forest model substantially outperformed linear regression in predicting wine quality, explaining approximately 54% of the variance compared to 40% for the linear model. Performance metrics demonstrated this clear advantage:

Model	RMSE	R^2	
Linear Regression	0.63	0.40	
Random Forest	0.55	0.54	

Feature importance analysis from the Random Forest model revealed that not all chemical properties contribute equally to wine quality (Figure 1). **Alcohol content** emerged as the dominant predictor (27% importance), followed by **volatile acidity** (11%), **sulphates** (14%), and **total sulfur dioxide** (7%).



The prediction accuracy visualization (Figure 2) shows that while the model performs well overall, it tends to underpredict quality for higher-rated wines and slightly overpredict for lower-rated wines. Most predictions fall within ±0.5 units of the actual rating, with better performance in the middle quality range (5-6) than at the extremes.



The superior performance of the Random Forest model suggests that the relationships between chemical properties and wine quality are complex and non-linear, which linear regression cannot fully capture.

Discussion

This analysis confirms that objective chemical measurements can significantly predict subjective quality assessments, with over half of the variation in expert ratings explained by physicochemical properties alone. These findings have several important implications for the wine industry.

The dominance of alcohol content as a predictor aligns with oenological understanding—higher alcohol is associated with greater ripeness of grapes, which typically contributes to fuller body and more intense flavors. However, this relationship isn't simply "more is better," as excessive alcohol can create imbalance. The negative relationship between volatile acidity and quality confirms the general understanding that acetic acid, which creates vinegar-like aromas, is considered a fault in most wines beyond minimal levels.

For winemakers, these results suggest several practical strategies for quality improvement: managing fermentation conditions to minimize volatile acidity production, considering appropriate sulphate additions to prevent oxidation, and focusing on viticultural practices that promote optimal grape ripeness for balanced alcohol levels.

A significant limitation of this study is that it explains only about half of the quality variation, indicating that important factors remain unaccounted for. These likely include variables not captured in the dataset: grape variety, vineyard location, vintage conditions, winemaking techniques, and aging methods. Additionally, the dataset contained only Portuguese red wines, potentially limiting generalizability to other wine styles and regions.

The subjective nature of quality assessment itself presents another limitation. Different expert panels might prioritize different aspects of wine, and cultural preferences vary globally. Future work should investigate whether these chemical-quality relationships hold across different wine styles, regions, and expert panels.

An intriguing expansion would be incorporating sensory descriptors alongside chemical measurements and quality ratings. This could illuminate which specific sensory attributes

(fruitiness, tannin structure, etc.) are influenced by which chemical properties, creating a more complete bridge between objective measurements and subjective experience.

Despite these limitations, this analysis demonstrates that wine quality is not purely subjective—there are measurable, objective properties that significantly influence expert perception. In an industry where art and science intersect, this research helps quantify that relationship, offering both practical insights for producers and deeper understanding for consumers.

Figures and Captions

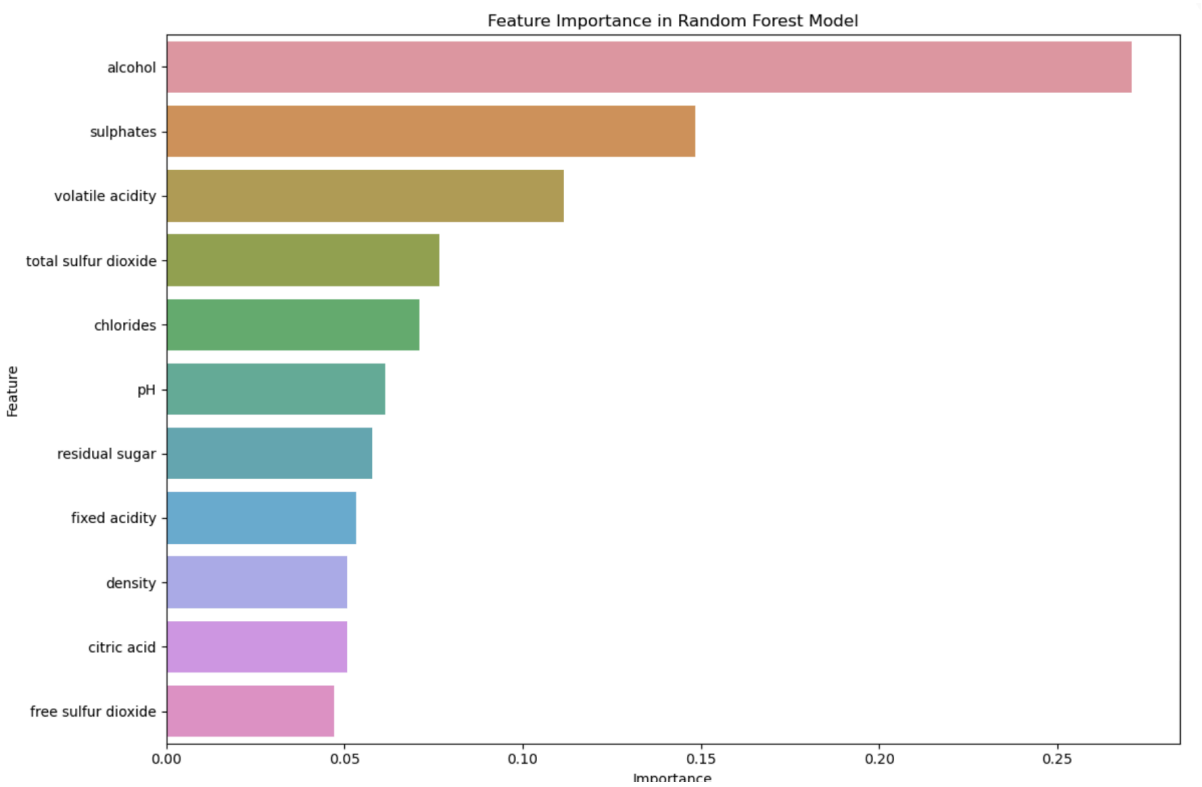


Figure 1: Feature Importance in Predicting Wine Quality

Relative importance of physicochemical properties in predicting wine quality from the Random Forest model. Alcohol content shows the highest importance (27%), followed by volatile acidity (11%) and sulphates (14%).

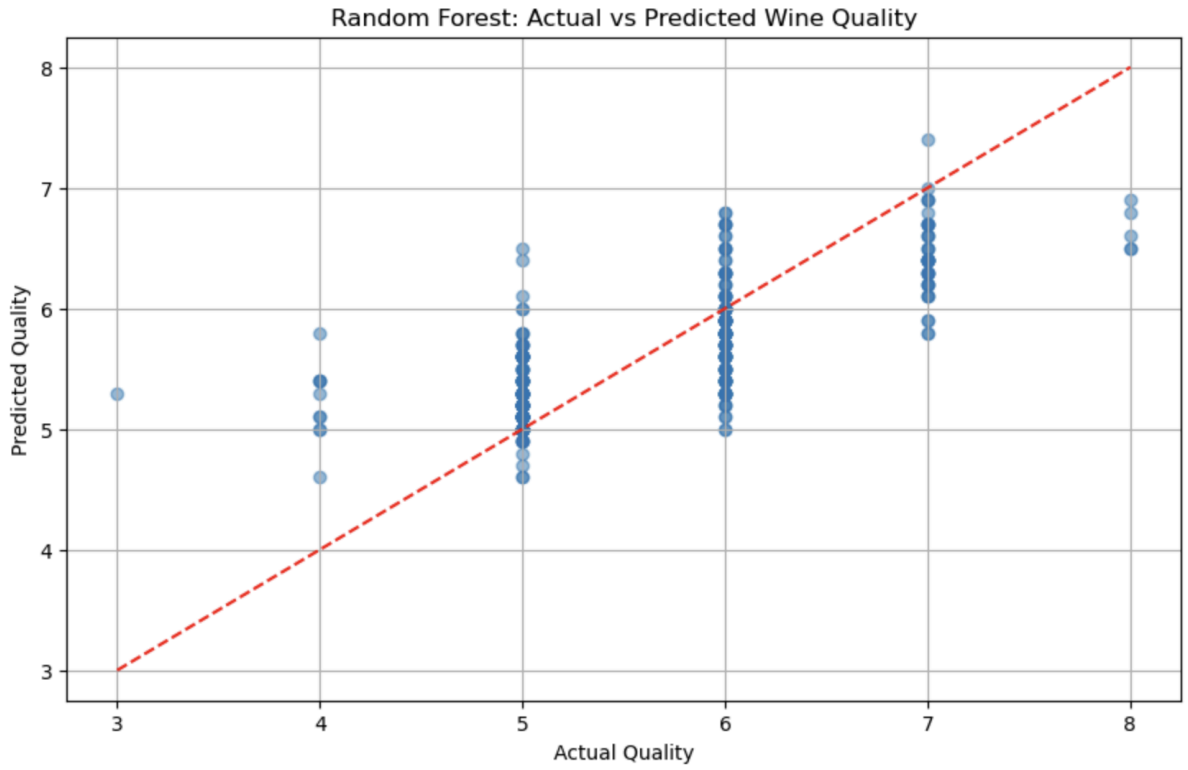


Figure 2: Actual vs. Predicted Wine Quality

Comparison of actual expert ratings with Random Forest model predictions. Most predictions fall within ±0.5 units of actual ratings, with better accuracy in the middle quality range (5-6). The diagonal line represents perfect prediction.

Citations and Attributions

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T
- Dataset URL: https://archive.ics.uci.edu/dataset/186/wine+quality
- Boulton, R., Singleton, V., Bisson, L., & Kunkee, R. (1996). *Principles and Practices of Winemaking*. Springer.