Predicting Wine Quality: A Binary Classification Approach Using Physicochemical Properties

Table of contents

Introduction 1
- Research Question . . . 2
- Dataset Description . . . 2

Methods 2

Exploratory Data Analysis 3

Model Development and Training 4

Results 4

Summary 5

Discussion 5

References 6

Authors: Aidan Hew, Karan Bains, Shuhang Li

Introduction

Wine quality assessment is traditionally performed by human experts through sensory evaluation, a process that is subjective, time-consuming, and requires specialized training. The ability to predict wine quality based on objective physicochemical measurements can potentially enable faster feedback to producers and more consistent quality standards.

Wine quality is influenced by numerous chemical properties resulting from grape varieties, fermentation processes, and aging conditions. Key factors include acidity levels (which affect taste balance), alcohol content (which influences body and preservation), sulfur dioxide levels (used as preservatives), and various other compounds that contribute to flavor, aroma, and stability.

Research Question

Can we accurately predict whether a red wine is of high quality based solely on its physicochemical properties?

We aim to build a binary classification model that distinguishes between high-quality wines (rated 7 out of 10) and lower-quality wines (rated <7) using features such as acidity, alcohol content, sulfur dioxide levels, and other measurable chemical properties.

Dataset Description

This analysis uses the Red Wine Quality dataset, which is publicly available from the UCI Machine Learning Repository. The dataset contains 1599 samples of red Portuguese "Vinho Verde" wine, collected between 2004-2007. Each wine sample includes 11 physicochemical features and a quality score.

Features: - Fixed acidity (g/L): Non-volatile acids (tartaric acid) - Volatile acidity (g/L): Acetic acid content (high levels indicate vinegar taste) - Citric acid (g/L): Adds freshness and flavor - Residual sugar (g/L): Sugar remaining after fermentation - Chlorides (g/L): Salt content - Free sulfur dioxide (mg/L): Prevents microbial growth - Total sulfur dioxide (mg/L): Total SO (bound and free) - Density (g/cm³): Wine density - pH: Acidity level (scale 0-14) - Sulphates (g/L): Wine additive contributing to SO - Alcohol (% volume): Alcohol content

Target Variable: - Quality: Expert ratings on a scale from 0 (very bad) to 10 (excellent)

Methods

We trained three classification models; Logistic Regression, Decision Tree, and Random Forest, to predict high versus low quality wines. The data was split 80/20 with stratification, with all features being standardised before model training. Due to the under-representation of high-quality wines in our data, we trained all models with balanced class weights to compensate for this limitation. Model performance was then evaluated using accuracy, precision, recall, F1 score, and ROC AUC, with the best performing model undergoing 5-fold cross-validation.

Exploratory Data Analysis

Unnamed: 0	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	973	973	973	973	973	973	973	973	973	973	973
mean	8.239	0.528	0.256	2.274	0.078	15.767	45.423	0.997	3.321	0.64	10.437
std	1.683	0.18	0.188	0.608	0.016	9.896	30.877	0.002	0.149	0.134	1.066
min	4.6	0.12	0	0.9	0.012	1	6	0.99	2.86	0.33	8.4
25%	7.1	0.39	0.08	1.9	0.068	8	23	0.996	3.22	0.55	9.5
50%	7.9	0.52	0.24	2.2	0.078	14	37	0.997	3.32	0.61	10.1
75%	9.1	0.64	0.4	2.5	0.087	21	60	0.998	3.41	0.71	11.1
max	15	1.33	0.76	4.7	0.147	57	165	1.002	3.9	1.22	14

Table 1: Summary statistics for physicochemical features

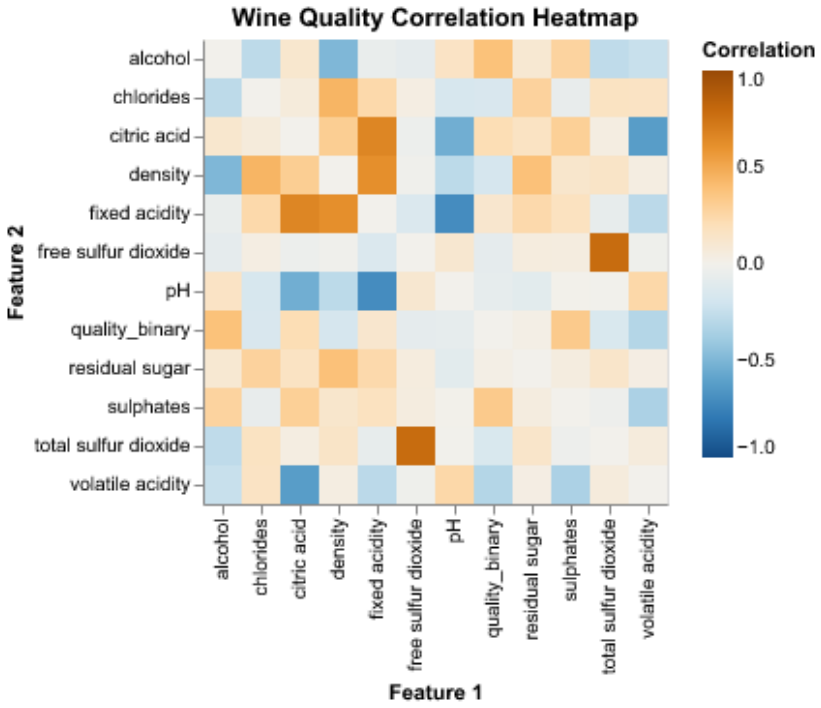

Figure 1: Correlation among features.

Figure 1 showed only a few features that displayed meaningful relationships with the quality of wine. In particular, alcohol, sulphates, and volatile acidity exhibited the strongest correlations

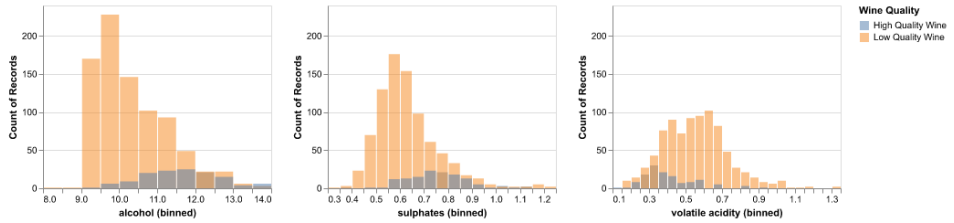

Figure 2: Distributions of features by wine quality

with our target. High-quality wines were associated with higher alcohol and sulphate levels but lower volatile acidity. The Figure 2 helped to demonstrate this separation in distribution of high and low-quality wines for these features. Finally, the dataset contained a notable imbalance, with relatively few high-quality wines, implying the need for class balancing in the later modelling stages.

Model Development and Training

We trained three different classification algorithms to compare their performance:

1. **Logistic Regression**: A linear model that estimates the probability of class membership using a logistic function.
2. **Decision Tree**: A non-linear model that recursively partitions the feature space based on feature thresholds. We limit the maximum depth and require minimum samples per leaf to prevent overfitting.
3. **Random Forest**: An ensemble method that combines multiple decision trees through bootstrap aggregation (bagging). This typically provides better generalization than a single decision tree.

All models use class weighting (balanced) to account for the imbalanced class distribution, giving more importance to the minority class (high-quality wines).

Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.807811	0.782787	0.364865	0.818182	0.504673	0.855378
Decision Tree	0.84481	0.754098	0.329114	0.787879	0.464286	0.818469

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	0.945529	0.868852	0.512195	0.636364	0.567568	0.845613

Table 2: Performance metrics for each evaluated model

The Random Forest Classifier performed the strongest across all metrics, achieving 0.869 test accuracy as well as an ROC AUC of 0.846. Cross-validation confirmed this result, with a mean accuracy of 0.874 and relatively low variance across folds. This suggests that the model will generalise well, and indicates potential non-linear relationships in the data due to its superiority over the Logistic Regression and Decision Tree models.

Summary

This analysis investigates whether physicochemical properties (eg. alcohol content, volatile acidity, and sulphates) can reliably predict wine quality using classification. Using a dataset of 1599 red Portuguese "Vinho Verde" wines (Cortez et al. 2009b), we developed models to distinguish between high-quality wines (rated 7 or higher) and lower-quality wines (rated below 7). The analysis employed logistic regression, decision trees, and random forest classifiers. Results indicate that alcohol content, volatile acidity, and sulphates are the strongest predictors of wine quality, with the random forest model achieving 0.869 accuracy and an AUC of 0.846. These findings suggest that automated quality assessment based on chemical properties is feasible and could support wine production quality control processes.

Discussion

Our results indicate that a small subset of physiochemical properties carry most of the predictive ability in the how the quality of wines are perceived. Intuitively, volatile acidity was negatively correlated to the quality of wine. Surprisingly, alcohol was positively correlated with our target, defying theory which suggests that high levels of alcohol can reduce the aromas of wine, thus making it less enjoyable to consumers (Ozturk and Anli 2014). Further defying established research, sulphates were also positively correlated with wine quality, though modern techniques usually render these to have little effect on the flavour and smell of wine (Bakker et al. 1998). The strong performance of the Random Forest Classifier on this data in comparison to the Logistic Regression and Decision Tree Classifier seems to suggest some non-linear relationships within the data.

While the Random Forest Classifier obtained relatively impressive predictive ability on the test and cross-validation sets, there are several key limitations to the model. The dataset

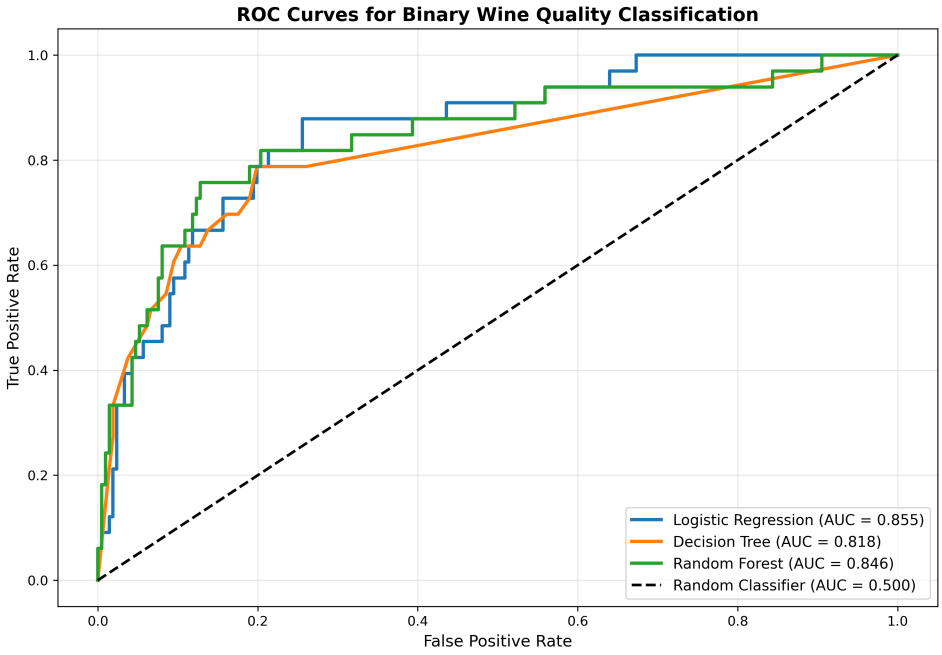

Figure 3: ROC Curves for Binary Wine Quality Classification

contained just under 1599 observations, with each of these belonging to the same grape and region. For these reasons, it is unlikely that the model would generalise well on a dataset containing a variety of red wines. Additionally, while the quality evaluations were produced by blind tastings from expert oenologists (Cortez et al. 2009a), these results are still highly subjective. Given these limitations, in future, the model must be validated on other external datasets to validate its robustness in predicting high quality red wines.

References

Bakker, J., P. Bridle, C. Garcia-Viguera, and et al. 1998. "Effect of Sulphur Dioxide and Must Extraction on Colour, Phenolic Composition and Sensory Quality of Red Table Wine." *Journal of the Science of Food and Agriculture* 78 (3): 297–307. https://doi.org/10.1002/(SICI)1097-0010(199811)78:3%3C297::AID-JSFA117%3E3.0.CO;2-G.

Cortez, Paulo, Luis Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009a. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547–53. https://doi.org/10.1016/j.dss.2009.05.016.

———. 2009b. "Wine Quality." UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Ozturk, Burcu, and Ertan Anli. 2014. "Different Techniques for Reducing Alcohol Levels in Wine: A Review." *BIO Web of Conferences* 3. https://doi.org/10.1051/bioconf/

20140302012.