Classification of Red Wine Quality Using Logistic Regression, SVM, and Random Forest

Chen, Aaron
aaronc@wustl.edu

Dong, Thomas
d.liangtong@wustl.edu

Zhao, Yipeng
godcallray@gmail.com

***Abstract*—This paper presents an exploratory analysis and comparative study of machine learning models for classifying red wine quality. The dataset comprises 1599 samples with 11 physicochemical features. Initial exploratory data analysis revealed that wine quality scores predominantly fall into the medium range (5-6), with alcohol content showing the strongest positive correlation and volatile acidity the strongest negative correlation with quality. Physicochemical features were preprocessed, and quality labels were binned into three categories: Low, Medium, and High. Three primary supervised learning models – Logistic Regression, Support Vector Machine (SVM), and Random Forest – were trained and evaluated. Experimental results show that Random Forest achieved the highest accuracy (0.8719), while SVM yielded the best macro F1-score (0.5691), indicating a better balance in classifying across the different quality categories. The study discusses the performance nuances of each model and suggests avenues for future research, including advanced feature engineering and alternative modeling techniques to improve classification, particularly for minority classes.**

***Index Terms*—wine quality, classification, machine learning, exploratory data analysis, feature correlation, logistic regression, support vector machine, random forest, data preprocessing, feature scaling, accuracy, F1-score, confusion matrix**

I. Introduction

Machine learning techniques are increasingly pivotal in various industries for predictive analytics, and the food and beverage sector, particularly oenology, is no exception. In this study, machine learning algorithms are applied to evaluate and predict the quality of red wine based on its objectively measurable physicochemical properties. Predicting wine quality is a valuable task for the wine industry, aiding in quality control, product development, and market positioning. This paper focuses on developing and comparing different classification models to predict wine quality as a categorical variable. This study utilizes a dataset of 1599 red wines, each characterized by 11 features. To facilitate classification, wine quality scores were binned into "Low," "Medium," and "High" categories. This paper compares the performance of Logistic Regression, Support Vector Machine (SVM), and Random Forest models in predicting these quality categories, following data preprocessing steps including feature scaling. The performance of these models in classifying wine quality into the defined categories is rigorously evaluated using several standard metrics, including overall accuracy, the macro-averaged F1-score (to account for potential class imbalance in the binned categories), and detailed confusion matrices to analyze the specific prediction patterns for each quality class.

II. Exploratory Analysis of the Data Set

A. Dataset Overview

The dataset employed for this study is the 'Red Wine Quality' dataset, a well-regarded collection sourced from the UCI Machine Learning Repository. It provides a comprehensive set of attributes based on physicochemical tests conducted on samples of Portuguese "Vinho Verde" wine. The input variables are all objective measurements, providing a quantitative foundation for assessing wine quality.

After loading the `winequality-red.csv` file, an initial check using `df.info()` and `df.head()` reveals 1599 valid data points, each described by 11 distinct physicochemical features. These features are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output variable for this dataset is `quality`, which represents a sensory score assigned by experts. In this specific dataset, these scores range from 3 to 8, where higher values indicate a higher perceived wine quality.

We first use a countplot of quality to show a distribution of our output data. We see that the distribution of wine quality scores is not so uniformed, with most wines having a quality rating between 5 and 6. Higher quality scores (7, 8) are rare, and extremely low scores are even more rare.

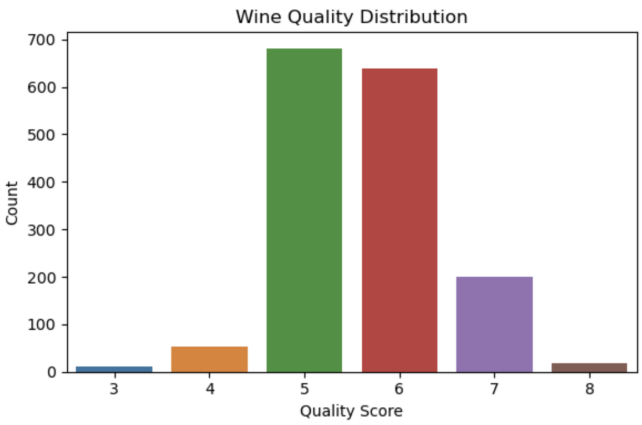

Fig. 1. Overview of Quality Distribution

B. Feature Correlation

We use a heatmap to display the correlation between each variables. The correlation heatmap reveals several important

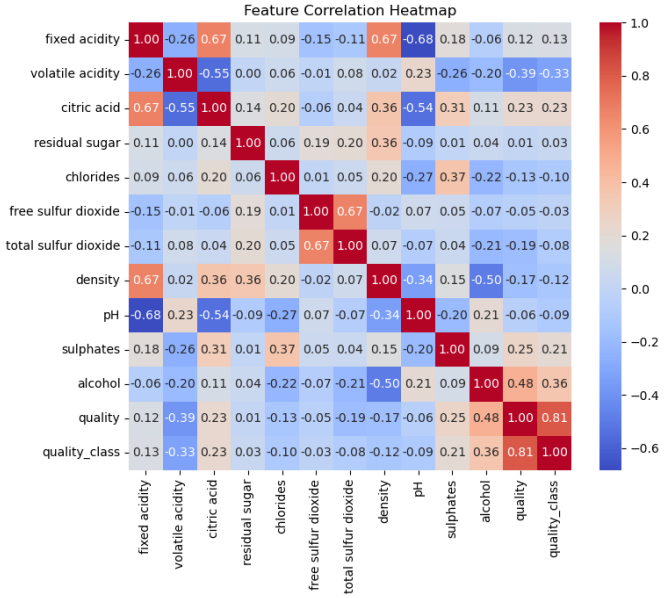

Fig. 2. Heatmap Feature Correlation

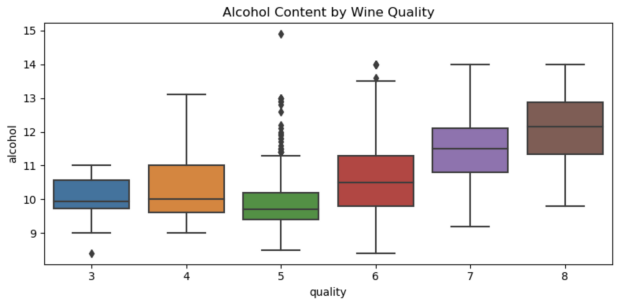

Fig. 3. Alcohol vs. Quality BoxPlot

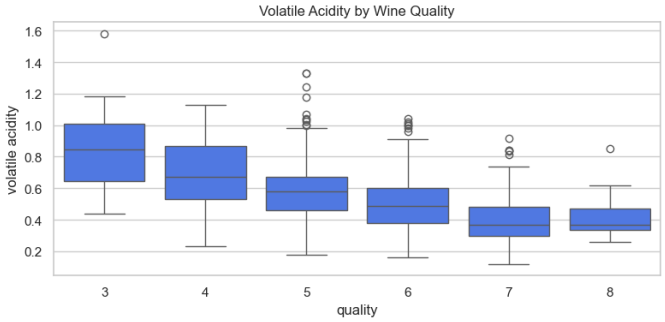

Fig. 4. Alcohol vs. Quality BoxPlot

relationships between the features and the target variable, `quality`. Notably, no single feature exhibits a strong linear correlation with `quality`. In particular, features such as `residual sugar` and `pH` show correlations close to zero, suggesting that their relationship with wine quality may be nonlinear. This observation highlights the potential need for feature normalization or the use of models capable of capturing nonlinear patterns.

Among all features, `alcohol` shows the strongest positive correlation with wine quality, while `volatile acidity` demonstrates the strongest negative correlation. These two features are likely to be more influential predictors and may gain greater emphasis in model training, either through feature weighting or selection.

Noticeably, some features are highly correlated with each other. For example, `free sulfur dioxide` with `total sulfur dioxide`, and `density` with `residual sugar`.

To gain deeper insights, we will next examine the features most strongly associated with quality in greater detail.

C. Alcohol vs. Quality

The boxplot shown in Figure 3 indeed demonstrates a positive relationship. Specifically, wines with a quality rating of 3 to 5 have lower median alcohol levels (around 10%), while those rated 7 or 8 exhibit noticeably higher median alcohol levels (around 11.5% to 12%). Such significant difference in median values suggests that `alcohol` could be a significant distinguisher between high-quality and low-quality wines.

D. Volatile Acidity vs. Quality

The boxplot shown in Figure 4 indeed demonstrates a positive relationship. Specifically, wines with a quality rating of 3 to 5 have lower median alcohol levels (around 10%), while those rated 7 or 8 exhibit noticeably higher median alcohol levels (around 11.5% to 12%). Such significant difference in median values suggests that `alcohol` could be a significant distinguisher between high-quality and low-quality wines.

III. Modeling Methods

Our approach to modeling the wine quality classification problem is structured in four main phases: data preprocessing, label binning, model training, and evaluation. The pipeline is built in Python using the scikit-learn library.

A. Data Preparation

We begin by loading the `winequality-red.csv` dataset, which includes 11 physicochemical features and a quality score ranging from 3 to 8. The data is converted from semicolon-separated text to numerical format using `pandas`. Missing values are not present in this dataset.

B. Label Encoding and Binning

To formulate the classification task, we map the numerical quality scores into three categories:

- Low (quality ≤ 4)
- Medium (quality 5 or 6)
- High (quality ≥ 7)

This transformation addresses the original class imbalance and makes the classification more robust and interpretable.

C. Feature Scaling

Since SVM and logistic regression are sensitive to feature magnitudes, we apply standardization to all input features using `StandardScaler`. Random forest, being a tree-based model, is insensitive to scaling but uses the same scaled data for consistency.

D. Model Training

We use three supervised learning models:

- Logistic Regression: Configured with `multi_class='multinomial'`, `solver='lbfgs'`, `max_iter=1000`, and `class_weight='balanced'`.
- Support Vector Machine (SVM): Configured with an RBF kernel, `C=1.0`, `gamma='scale'`, and `class_weight='balanced'`.
- Random Forest: A range of models are trained with increasing `n_estimators` and fixed `max_depth=10`, using `class_weight='balanced'`.

Each model is trained using an 80/20 train-test split, stratified by class label to preserve class distribution.

E. Evaluation Metrics

We evaluate the models based on:

- Accuracy: Overall proportion of correct predictions.
- Macro F1-score: Harmonic mean of precision and recall computed independently for each class and then averaged.
- Confusion Matrix: For visualizing prediction distribution across classes.

IV. Experimental Results and Analysis

To evaluate model performance, we compared Logistic Regression, Support Vector Machine (SVM), and Random Forest classifiers using accuracy and macro-averaged F1-score.

A. Performance Summary

The performance comparison is summarized in Table I, showing that Random Forest achieved the highest accuracy, while SVM had the best macro F1-score.

TABLE I
Model performance comparison on the test set

Model	Accuracy	Macro F1-score
Logistic Regression	0.6156	0.5035
Support Vector Machine	0.7063	0.5691
Random Forest	0.8719	0.5327

B. Confusion Matrix Visualization

We generated confusion matrices for all three classifiers. Each matrix visualizes the predicted versus true labels, revealing strengths and weaknesses in classifying "Low", "Medium", and "High" quality wines. These are illustrated in following figures.

Fig. 5. Confusion Matrix of Logistic Regression

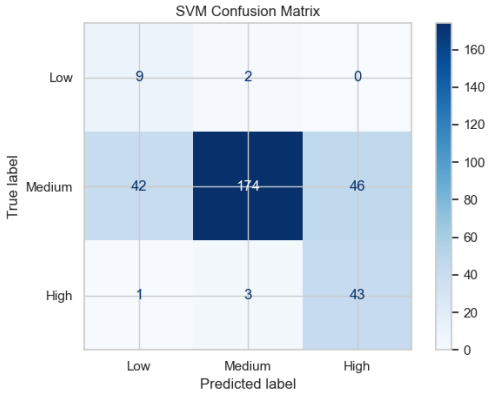

Fig. 6. Confusion Matrix of SVM

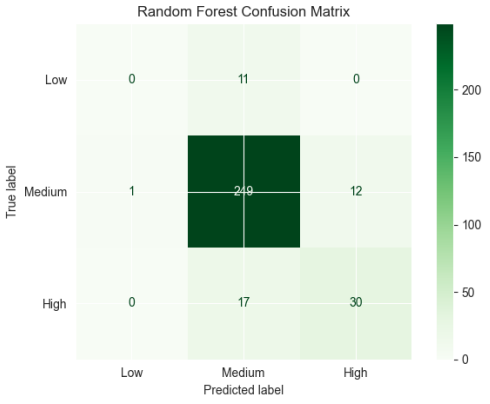

Fig. 7. Confusion Matrix of Random Forest

C. Performance Trends with Varying Estimators

To further understand how Random Forest behaves with different numbers of trees, we plotted accuracy and macro F1-score over varying `n_estimators`. As shown in Figure 8, performance tends to stabilize after 300 trees, suggesting diminishing returns with additional complexity.

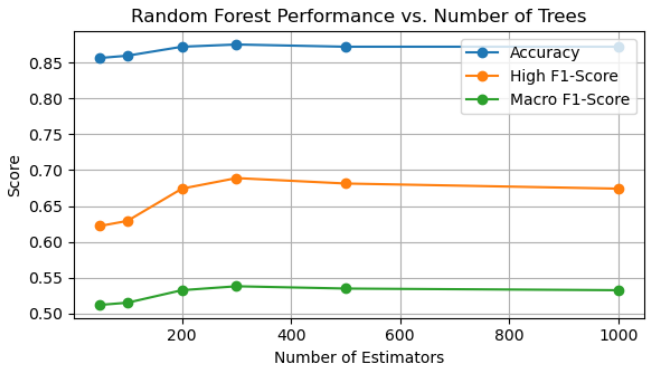

Fig. 8. Effect of `n_estimators` on Random Forest Accuracy and F1

V. Conclusions

This study compared four machine learning classifiers for red wine quality prediction, with detailed results presented for Logistic Regression, Random Forest, and SVM. After categorizing wine quality into Low, Medium, and High, Random Forest achieved the highest overall accuracy (0.8719), while the Support Vector Machine yielded the best macro-averaged F1-score (0.5691), indicating a better balance across classes. These findings highlight that Random Forest excels in general classification, whereas SVM is more adept at equitably identifying different quality tiers. Future work could explore advanced feature selection, deeper neural architectures, and cost-sensitive learning to further improve predictions, especially for minority classes.

Author Contributions: Aaron Chen performed ; Thomas Dong implemented ; Yipeng Zhao built .

Appendix