Red Wine Quality Prediction Model

Group 6: Ziming Qin, Yiwen Ma, Lok LamWong, Shangzhou Xia

I. Introduction

In the competitive and nuanced world of wine sales, the ability to accurately predict wine quality can serve as a crucial edge for businesses. Our initiative focuses on leveraging machine learning models to assess and predict wine quality with high precision. By employing decision tree, logistic regression, and random forest models, we have embarked on a data-driven approach to discern the subtleties that distinguish high-quality wines from the rest. This method is not only cost-effective but also resource-efficient, making it ideal for smaller companies looking to gain a competitive advantage in the market.

II. Business Objectives

Our primary objective is to develop a robust system that can accurately predict the quality of wine. This system will serve three main stakeholders: suppliers, customers, and retailers, by fulfilling the following business objectives:

For suppliers, enable wine suppliers to assess the quality of their products before distribution. This predictive capability ensures that only wines of a certain quality standard are introduced into the market, thereby enhancing the supplier's reputation and product value. For customers, providing customers with a reliable indication of wine quality helps them make informed purchasing decisions. By leveraging our predictive models, customers can have confidence in the quality of the wine they purchase, improving their overall satisfaction and trust in retailers. For retailer partnership, working with retailers to incorporate our predictive models into their inventory selection and marketing strategies. Retailers can use quality predictions to curate a selection of wines that meets the preferences and expectations of their customers, optimizing stock levels and reducing the risk of unsold inventory.

III. Results

A. *Wine quality prediction*

To assess and predict wine quality, we employed the decision tree, logistic regression, and random forest models because they are relatively cost-effective and resource-efficient for smaller companies. After fine-tuning their hyperparameters, we compared their performance on the test set by measuring accuracy, precision, recall, F1 score, and the Area under the ROC Curve (AUC).

	Decision Tree	Logistic Regression	Random Forest
Accuracy	0.756	0.747	0.822
Precision	0.773	0.769	0.836
Recall	0.773	0.756	0.831
F1 score	0.773	0.762	0.834
AUC	0.80	0.83	0.91

Table 1. Performance Metrics of the 3 models

Table 1 shows that the random forest model achieved the best performance, with the highest accuracy (0.822), precision (0.836), recall (0.831), and F1-score (0.834). Meanwhile, the decision tree and logistic regression models have similar performances but are worse than the random forest model. It reveals that the random forest model has the best predictive capability for wine quality, which helps us identify high-quality wines for sales or other business strategies.

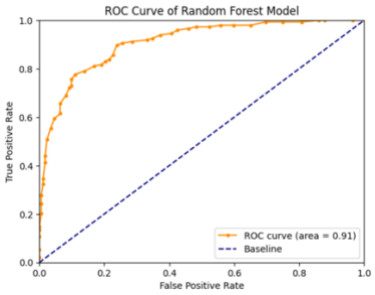
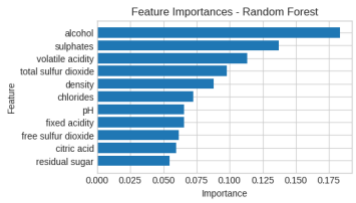


Figure 1. Receiver Operating Characteristic curve of Random Forest Model

Table 1 and Figure 1 also indicate that the random forest model has the best performance on the AUC (0.91), surpassing the decision tree (0.80) and logistic regression models (0.83). The ROC curve in Figure 1 highlights that the random forest model has a higher true positive rate in predicting high-quality wines while maintaining a lower false positive rate for different decision thresholds. It implies that the random forest model can better learn the patterns between wine characteristics and its quality, thereby correctly predicting the wine quality.

B. *Feature Importance*




```
Feature Importances:
                 Feature  Importance
10               alcohol    0.183361
9              sulphates    0.137153
1       volatile acidity    0.113269
6    total sulfur dioxide    0.098153
7                density    0.088177
4              chlorides    0.072543
8                     pH    0.065732
0          fixed acidity    0.065538
5    free sulfur dioxide    0.061832
2           citric acid    0.059473
3        residual sugar    0.054770
```

Fig 2. Feature importance of the model.

A score was calculated for each feature in our Random Forest model to find important features and provide better suggestions. As shown in Fig 2, the score simply represents the importance of each feature and the top 3 features are Alcohol (18.34%), Sulphates (13.72%), and Volatile Acidity (11.33%). However, we cannot know the direction between the feature and the target based on it.