RIT

Rochester Institute of Technology

MGIS 650 FINAL PREDICTION PROJECT

EXECUTIVE SUMMARY

Wine Quality Prediction with R

Ba Long Dang	bd8923@rit.edu
Seshabhatter Aditya	as2797@rit.edu
Kumar Divya Bala	dk9114@rit.edu
Elandassery Rajeev Akshay	ae4285@rit.edu
Chinta Suhas	vc2023@rit.edu
Varaganti Basanth Kumar	bv8946@rit.edu

Table of Contents

I. Introduction

- The wine industry's global value exceeds $340 billion, making quality prediction a crucial economic factor. Our research aims to bridge the gap between traditional sensory evaluation and objective chemical analysis through advanced predictive modelling.
- This project focuses on predicting wine quality scores using advanced statistical and machine learning tools. By leveraging the WineQT dataset, we employed Exploratory Data Analysis (EDA), Random Forest (RF), and Support Vector Machines (SVM) to forecast wine quality based on various physicochemical attributes. This analysis aims to provide actionable insights into the factors influencing wine quality and develop a robust predictive model for future wine quality scores over the next two years.

II. Data Description

- The dataset used contains 1,141 rows, representing data from the past 12 years, with each year consisting of 85 rows to capture the most frequent combinations of features. The machine learning models will be trained on this data to predict results for the next two years, which corresponds to approximately 190 data points.
- The dataset comprises measurements of 12 variables for numerous wine samples:
  - Fixed acidity
  - Volatile acidity
  - Citric acid
  - Residual sugar
  - Chlorides
  - Free sulfur dioxide
  - Total sulfur dioxide
  - Density
  - pH
  - Sulphates
  - Alcohol
  - Quality (target variable)
- Quality is rated on a scale from 0 to 10, with higher values indicating better quality.
- Even though future wine quality can be predicted using current data, this approach is limited in accuracy as it does not account for changes in consumer preferences or market dynamics. To address this, we will predict future trends for each physicochemical feature individually and then use these predicted values to forecast wine quality.

III. Data Analysis

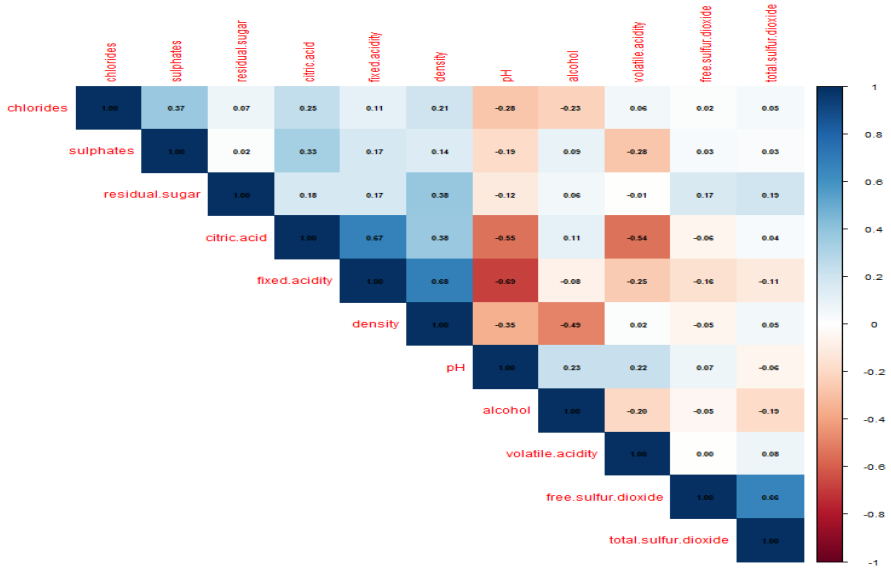
1. Wine Quality Distribution

- The distribution of wine quality scores follows a bell-shaped curve, ranging from 3 to 8, with most wines clustering around the middle scores.
- Scores of 5 and 6 are the most common, with 483 wines scoring 5 and 462 wines scoring 6, making up the bulk of the dataset.
- At the lower and higher ends, the number of wines is much smaller:
  - Only 6 wines received the lowest score of 3.
  - 33 wines scored 4.
  - 143 wines scored 7.
  - Just 16 wines scored 8, the highest possible score.
- This distribution shows a slight right skew, with more wines clustered toward the middle and a longer tail toward the higher scores. While exceptional quality (score of 8) is rare, it is still more common than very poor wines.
- The scarcity of wines in the lower range (scores 3-4) suggests that wines of poor quality are uncommon in this dataset. In contrast, the rarity of high scores highlights the difficulty of achieving exceptional quality.
- Overall, the distribution indicates that most wines maintain average quality, but reaching the highest level of excellence is a significant challenge.

2. Correlation Analysis

- The correlation matrix heatmap reveals intricate relationships between various wine quality factors, highlighting several key patterns:

- Strong positive correlations include:
  - Fixed acidity and citric acid (0.67).
  - Free sulfur dioxide and total sulfur dioxide (0.66).
- Significant negative correlations are found between:
  - pH and fixed acidity (-0.69).
  - pH and citric acid (-0.55).
- Moderate negative correlations include:
  - Alcohol and volatile acidity (-0.20).
  - Density and alcohol (-0.49).



- Some characteristics show weak or negligible correlations, such as:
  - Chlorides and pH (-0.28).
  - Residual sugar and alcohol (-0.01).
- The relationships between acidity factors (fixed acidity, citric acid, and pH) align with fundamental wine chemistry principles, while the strong correlation between free and total sulfur dioxide reflects their complementary roles in wine preservation.
- Alcohol's negative relationships with density and volatile acidity offer insights into its structural impact on wine. The negligible correlation between residual sugar and alcohol suggests these components vary independently in the winemaking process.
- This comprehensive analysis of chemical component relationships provides valuable insights into the factors influencing wine quality and characteristics.

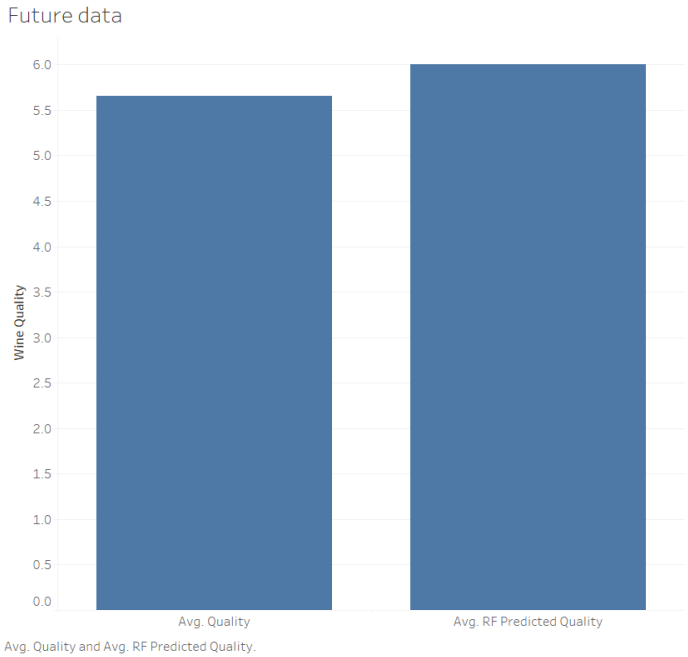
3. Training and choosing ML model

- Given the weak correlations among the features, all variables will be utilized in the model to ensure a comprehensive analysis that captures the unique contributions of each attribute to the overall prediction.
- Performance Metrics (check R file for result)

Model	Accuracy	Kappa
Random Forest	0.67	0.47
SVM	0.65	0.45

- Wine Quality Forecast
  - The average wine quality over the past 12 years is 5.8, while the predicted average for the next two years is projected to rise to 6.0.
  - This forecast indicates a notable improvement in average wine quality scores compared to the current dataset.
  - Machine learning predictions based on physicochemical properties highlight a trend toward producing higher-quality wines in the future.
  - These improvements likely reflect advancements in winemaking processes specific to this Portuguese wine, ensuring greater consistency in quality and consumer appeal.
  - The higher predicted quality also suggests:
    - A stronger market presence.
    - Increased consumer appreciation.

4. Impact on Sales and Pricing



- The predicted increase in wine quality suggests a potential rise in sales and pricing for the wine industry.
- Higher-quality wines typically attract greater consumer interest and justify premium pricing.
- With this forecast, producers can:
  - Anticipate increased demand.
  - Strategize effectively to maximize revenue.
- The focus on consistent quality improvements creates opportunities for:

- Enhanced brand loyalty.
- Greater market competitiveness.

❖ This analysis underscores the value of predictive modelling in:
- Driving profitability.
- Ensuring the long-term success of the wine industry.

IV. Conclusion and Recommendations

❖ The results of this analysis provide actionable insights into improving wine quality and guiding future production strategies. Key findings highlight the importance of optimizing specific physicochemical properties to enhance overall quality and consistency.

❖ Key Recommendations:

1. Optimize Acidity Profiles:
   - Focus on the strong positive correlation between fixed acidity and citric acid (0.67). Optimizing these components together can improve the wine's acidity balance, contributing to a more refined taste.
   - Adjust pH levels to complement acidity, as demonstrated by its significant negative correlation with fixed acidity (-0.69) and citric acid (-0.55).
2. Enhance Preservation Through Sulfur Dioxide Management:
   - Improve free sulfur dioxide and total sulfur dioxide levels, which show a strong positive correlation (0.66), to ensure better wine preservation and stability.
3. Refine Alcohol Content for Better Structure:
   - Address the moderate negative correlations between alcohol and volatile acidity (-0.20), as well as alcohol and density (-0.49). This can help achieve a better structural profile and enhance wine flavor.
4. Independent Optimization of Residual Sugar and Alcohol:
   - Given the negligible correlation (-0.01) between residual sugar and alcohol, these components should be adjusted separately to meet flavor and production goals.

❖ Strategic Implications:
- Implementing these adjustments can lead to consistent quality improvements, aligning with market demand for higher-quality wines.
- The predicted rise in average wine quality (from 5.8 to 6.0) provides an opportunity for producers to justify premium pricing and strengthen market competitiveness.
- Continuous use of predictive modeling can further refine production processes, ensuring long-term profitability and consumer satisfaction.

❖ By focusing on these key attributes and leveraging advanced machine learning techniques, wine producers can achieve greater quality consistency, enhance brand loyalty, and maintain a competitive edge in the global wine market.