Wine Quality Prediction

Aditya Seshabhatter (as2797)

Akshay Rajeev (ae4285)

Ba Long Dang (bd8923)

Basanth Kurmar Varaganti (bv8946)

Divya Bala Kumar (dk9114)

Suhas Chinta (vc2023)

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.

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:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. 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.

Data Analysis

Wine Quality Distribution

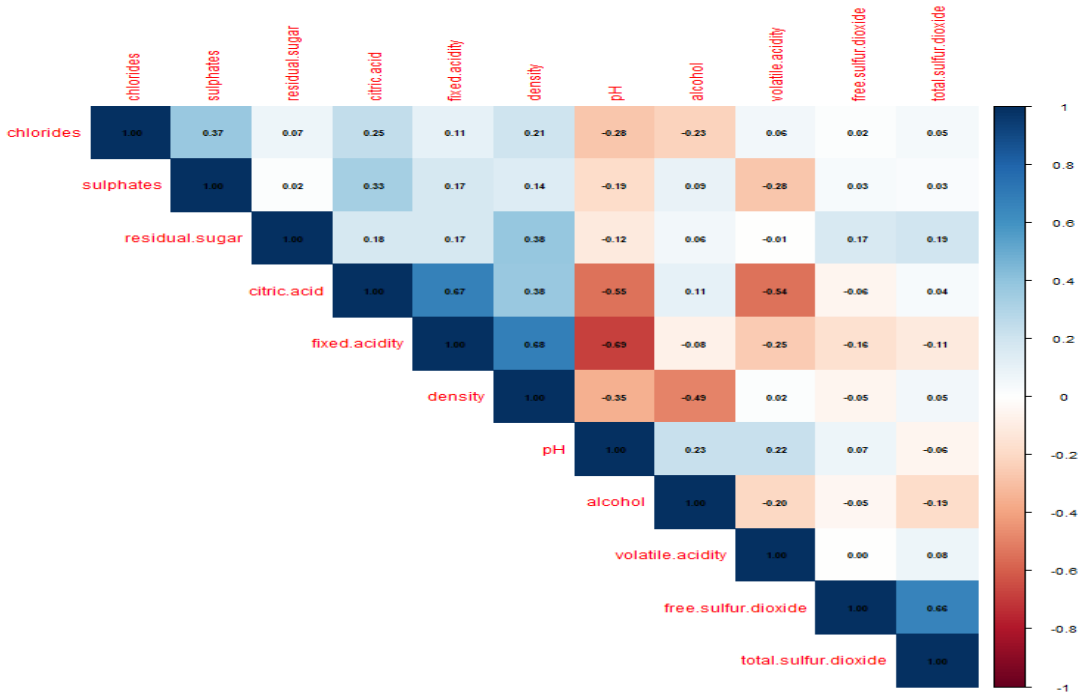
The distribution of wine quality scores follows a bell-shaped pattern ranging from 3 to 8, with most wines concentrated around the middle scores. Scores 5 and 6 dominate the dataset, with 483 wines scoring 5 and 462 scoring 6, representing the majority of the wines analyzed. At the extremes, the distribution reveals far fewer wines: only 6 wines received the lowest score of 3, 33 wines scored 4, 143 wines achieved a score of 7, and just 16 wines attained the highest score of 8. This pattern reflects a slight right skew, with a longer tail toward the higher scores, indicating that while achieving exceptional quality (score 8) is rare, it is more likely than producing truly poor wines. The scarcity of wines in the lowest quality range (scores 3-4) suggests that poor-quality wines are uncommon in this dataset. Conversely, the relative rarity of high scores highlights the challenge of producing wines of exceptional quality. Overall, the distribution underscores a general consistency in maintaining average quality standards while emphasizing the difficulty of reaching the highest levels of excellence.

Correlation Analysis

The correlation matrix heatmap reveals intricate relationships between various wine quality factors, with several notable patterns emerging from the data. Among the strongest positive correlations are those between fixed acidity and citric acid (0.67), as well as between free and total sulfur dioxide (0.66), indicating their closely linked roles in wine composition.

Significant negative correlations exist between pH and fixed acidity (-0.69), and pH and citric acid (-0.55), demonstrating the inverse relationship between acidity measures and pH levels.

The data also shows moderate correlations between alcohol and volatile acidity (-0.20), and between density and alcohol (-0.49), suggesting alcohol's complex influence on wine structure.



Some characteristics show weak or negligible correlations, such as chlorides and pH (-0.28) and residual sugar and alcohol (-0.01), indicating their independent variation. The interrelation of acidity factors (fixed acidity, citric acid, and pH) aligns with fundamental wine chemistry principles, while the strong correlation between free and total sulfur dioxide reflects their complementary preservation roles.

Alcohol's negative relationships with density and volatile acidity provide insights into its structural impact on wine, and the negligible correlation between residual sugar and alcohol suggests these components vary independently in wine production. This comprehensive analysis of chemical component relationships provides valuable understanding of the factors influencing wine quality and characteristics.

Data Analysis

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.

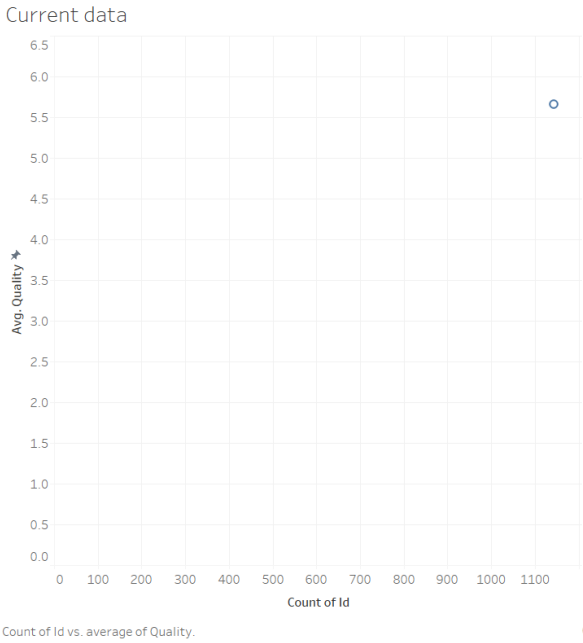
Performance Metrics (check R file for result)

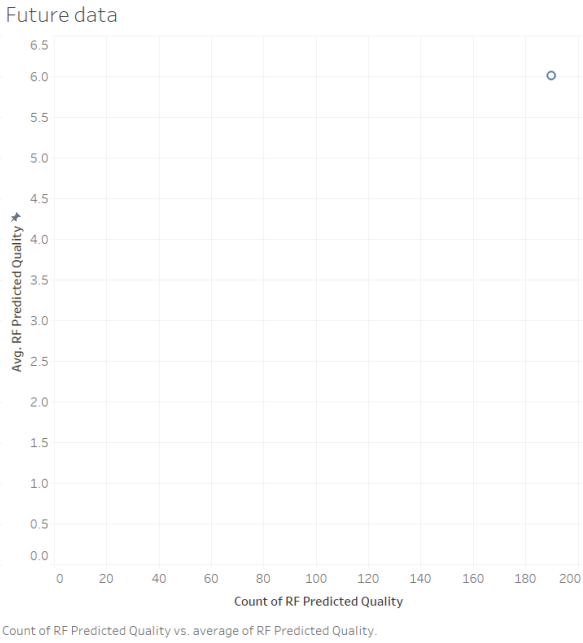
Model	Accuracy	Kappa
Random Forest	0.67	0.47
SVM	0.65	0.45

Given the better performance of the Random Forest model, we will use it for both predicting future wine features and leveraging these features to forecast future wine quality.

Wine Quality Forecast

The average wine quality over the past 12 years is 5.8. However, the predicted average wine quality for the next two years is projected to be 6.

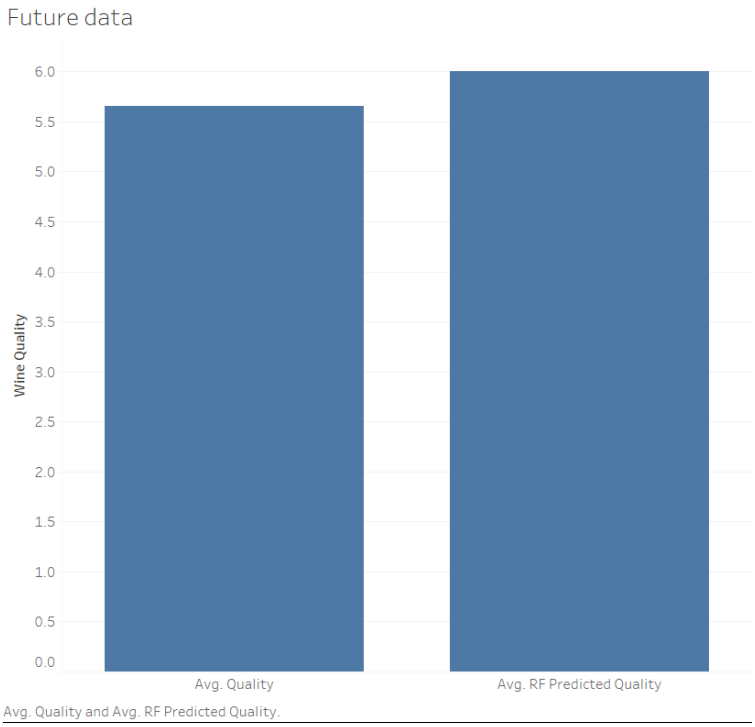




The forecasted quality for the wine industry indicates a notable improvement in average scores compared to the current dataset. Using machine learning predictions based on physicochemical properties, the model reveals a trend toward producing higher-quality wine in the future. This reflects potential improvements in winemaking processes specific to this Portuguese wine, ensuring consistent quality and appeal.

Higher predicted quality also suggests a stronger market presence and heightened consumer appreciation for this wine, positioning it favourably for future demand.

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 often attract greater consumer interest and justify premium pricing. With this forecast, producers can anticipate increased demand and strategize accordingly to maximize revenue. The focus on consistent quality improvements provides an opportunity for enhanced brand loyalty and market competitiveness. This analysis highlights the value of predictive modeling in driving profitability and ensuring the long-term success of wine.

Places of improvement

Based on the correlations in the dataset, the best features to focus on for improvement are fixed acidity, citric acid, sulfur dioxide levels, pH, alcohol content, and density.

The strong positive correlation between fixed acidity and citric acid (0.67) suggests that optimizing these acids together could enhance the wine's acidity profile. Additionally, the strong correlation between free and total sulfur dioxide (0.66) highlights the importance of improving sulfur dioxide levels for better preservation. Given the significant negative correlation between pH and both fixed acidity (-0.69) and citric acid (-0.55), adjusting the pH to a favorable range could improve the balance of acidity.

The moderate negative correlations between alcohol and volatile acidity (-0.20) and alcohol and density (-0.49) suggest that refining alcohol content could enhance the wine's flavor profile and structure. Finally, since residual sugar and alcohol show a negligible correlation (-0.01), these components should be optimized independently for desired flavor outcomes. Focusing on the interplay of these features will help improve the wine's overall quality and consistency.