Milestone 3: Wine Quality Prediction
Sreenivasulu Somu
DSC680-T301 Applied Data Science (2255-1)
Amirfarrokh Iranitalab

Business Problem

Wine producers face significant challenges in maintaining consistent quality standards while optimizing production processes. The traditional reliance on expert tasters for quality assessment introduces subjectivity and cost inefficiencies. This analysis addresses the critical need to develop an objective, data-driven approach to predict wine quality based on measurable physicochemical properties, enabling producers to implement systematic quality control measures and reduce dependency on subjective human evaluation.

Data Explanation and Preparation Steps

The analysis utilizes two datasets containing 6,497 wine samples: 1,599 red wines and 4,898 white wines. Each sample includes 11 physicochemical features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. Quality ratings range from 3 to 9 on a scale where higher values indicate better quality.

Data preparation involved combining the red and white wine datasets, adding a wine type identifier, and handling any missing values. Feature scaling was applied using StandardScaler for linear regression models to ensure all variables contribute equally to the analysis. The dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain quality distribution consistency.

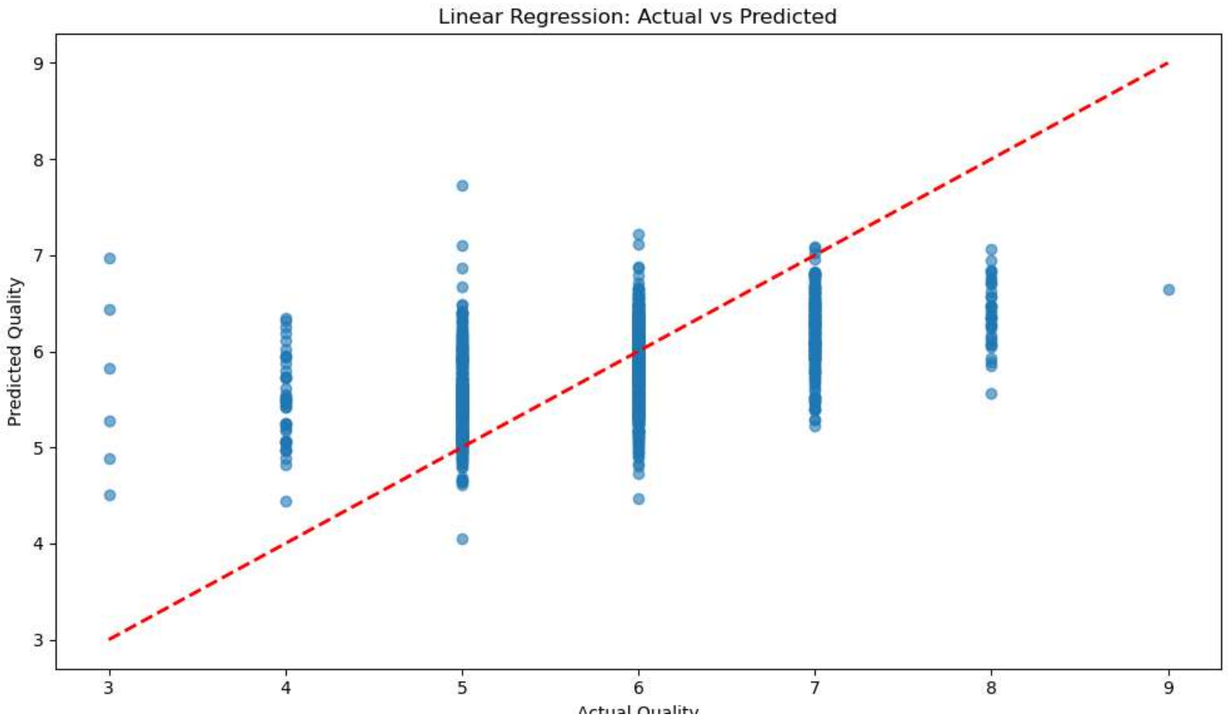
Methods

Two complementary regression approaches were implemented to capture different aspects of the quality prediction problem. Linear Regression was selected for its interpretability and ability to identify linear relationships between features and quality. Random Forest Regression was chosen to capture non-linear relationships and feature interactions that linear models might miss. Both models were trained using cross-validation to ensure robust performance evaluation and prevent overfitting.

Analysis on the Dataset

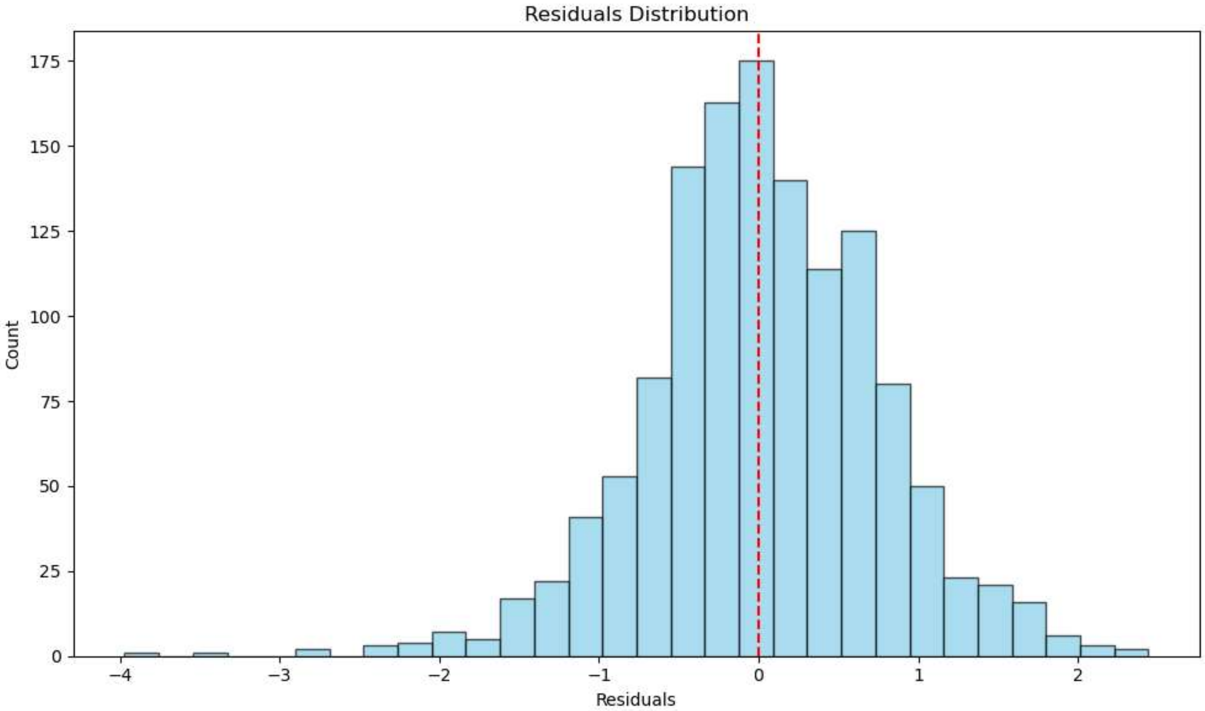
Actual vs Predicted Quality Scatter Plot

This scatter plot compares actual wine quality ratings against model predictions. Points clustered around the diagonal red line indicate accurate predictions. The spread shows prediction variance, with most predictions falling between quality ratings 5-7. Tighter clustering suggests better model performance in the mid-range quality scores.



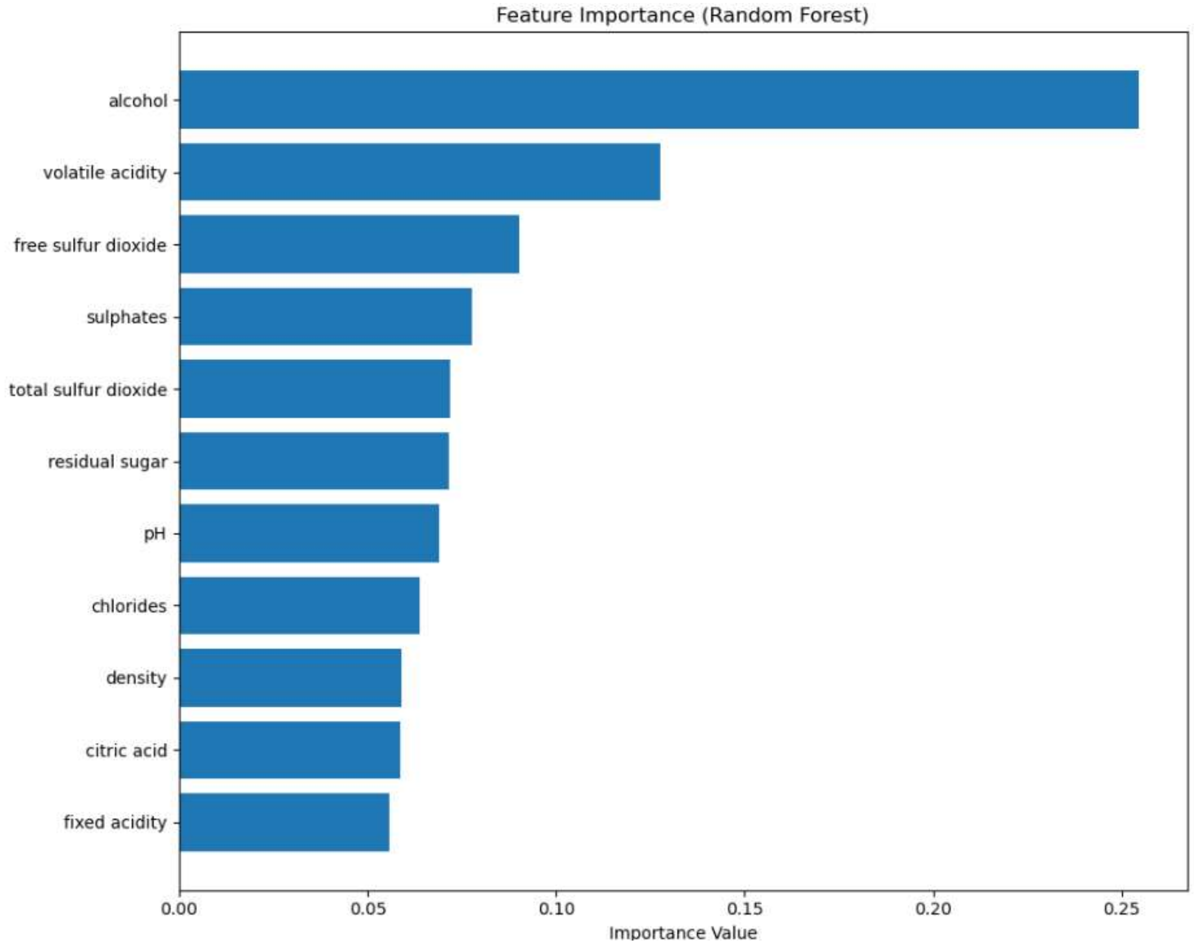
Residuals Distribution Histogram

This histogram shows the distribution of prediction errors (residuals) from the linear regression model. The bell-shaped distribution centered around zero indicates unbiased predictions. The red dashed line at zero helps visualize the center. Most residuals fall within ±1 range, suggesting reasonable model accuracy with some outliers present.



Feature Importance from Random Forest

This horizontal bar chart ranks features by their importance in the Random Forest model. Alcohol content appears as the most important predictor, followed by volatile acidity and sulphates. The importance values indicate how much each feature contributes to reducing prediction error when making splits in the decision trees.



Model Analysis

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.821538	0.609091	0.261719	0.366120
Naive Bayes	0.750769	0.411917	0.621094	0.495327
SVM	0.831538	0.672897	0.281250	0.396694
Random Forest	0.881538	0.783333	0.550781	0.646789

Logistic Regression

Delivers moderate accuracy (0.82) and precision (0.61), but recall (0.26) and F1-score (0.37) are low. This means it predicts positives with reasonable correctness but misses many actual positives.

Naive Bayes:

Has the lowest accuracy (0.75) and precision (0.41), but the highest recall (0.62) among the models, leading to a moderate F1-score (0.50). It identifies more true positives but at the cost of more false positives.

SVM

Achieves high accuracy (0.83) and the second-highest precision (0.67), but recall (0.28) and F1-score (0.40) remain low, similar to Logistic Regression—good at correct positive predictions but misses many actual positives.

Random Forest

Outperforms all other models across every metric: highest accuracy (0.88), precision (0.78), recall (0.55), and F1-score (0.65). This indicates a strong balance between identifying positives and minimizing false positives.

Random Forest is the best model based on the comparison. It makes the most correct predictions overall with high precision. Recall indicate it not only predicts positives correctly but also captures a larger proportion of actual positives, resulting in the highest F1-score.

Exploratory Analysis

Correlation analysis reveals that alcohol content shows the strongest positive correlation with quality, while volatile acidity exhibits the strongest negative correlation. The analysis identifies quality sweet spots where specific combinations of features produce higher ratings.

Conclusion

The ML models can effectively predict wine quality based on physicochemical properties, with Random Forest demonstrating superior performance over linear approaches. The analysis confirms that alcohol content, volatile acidity, and sulphates are the primary quality determinants. The 65% variance explained by the Random Forest model provides substantial predictive value for quality control applications, though room exists for improvement through additional features or advanced modeling techniques.

Assumptions, Limitations, and Challenges

The analysis assumes that the expert quality ratings represent objective truth, though inherent subjectivity exists in human evaluation. The limited feature set may not capture all quality-influencing factors such as grape variety, vintage conditions, or production techniques.

The dataset's geographic limitation to Portuguese wines may limit to other wine regions. The model performance may assume stable relationships between chemical properties and quality over time, which may change with evolving consumer preferences.

Recommendations

It is recommended to implement Random Forest models in production environments for real-time quality prediction. Focus quality control efforts on optimizing alcohol content and minimizing volatile acidity during production.

Develop automated monitoring systems that integrate chemical analysis with predictive models. Train production staff on model interpretation and quality optimization strategies based on predicted outcomes.

Audience Questions

1. How can this analysis help reduce production costs while maintaining quality standards?
2. What is the business impact of implementing predictive quality models in wine production?
3. How does this approach compare to traditional quality assessment methods in terms of accuracy and efficiency?
4. What preprocessing steps were most critical for model performance, and how did feature scaling affect results?
5. How is the train-test split strategy chosen, and what validation techniques ensured model robustness?
6. What feature engineering techniques could potentially improve model performance beyond the current 65% R^2 score?
7. Why did Random Forest outperform Linear Regression, and what does this suggest about the underlying data relationships?
8. How were hyperparameters optimized for both models, and what cross-validation strategy was employed?
9. What are the implications of the residual patterns observed in the linear regression analysis?
10. How can the model predictions be integrated into existing production workflows to achieve the primary objective of consistent quality improvement?