Name: Abdul Rehman
Student Number: 23113350
GitHub Repository: [https://github.com/AR-KASHMIRI47/REFERRAL-Clustering-and-Fitting]

Wine Quality Analysis Report

Introduction

This report explores the relationship between wine quality and its chemical properties using various data analysis techniques. The objective is to identify key factors influencing wine quality and develop models for classification and prediction. The analysis focuses on clustering wine samples based on chemical properties and predicting wine quality using regression models.

Dataset Description: Wine Quality Data (Kaggle - Ghassen Khaled)

This dataset consists of red and white wine samples from the Vinho Verde region of Portugal, characterized by several physicochemical properties and a quality rating assigned by wine experts.

Dataset Link: Kaggle Wine Quality Data

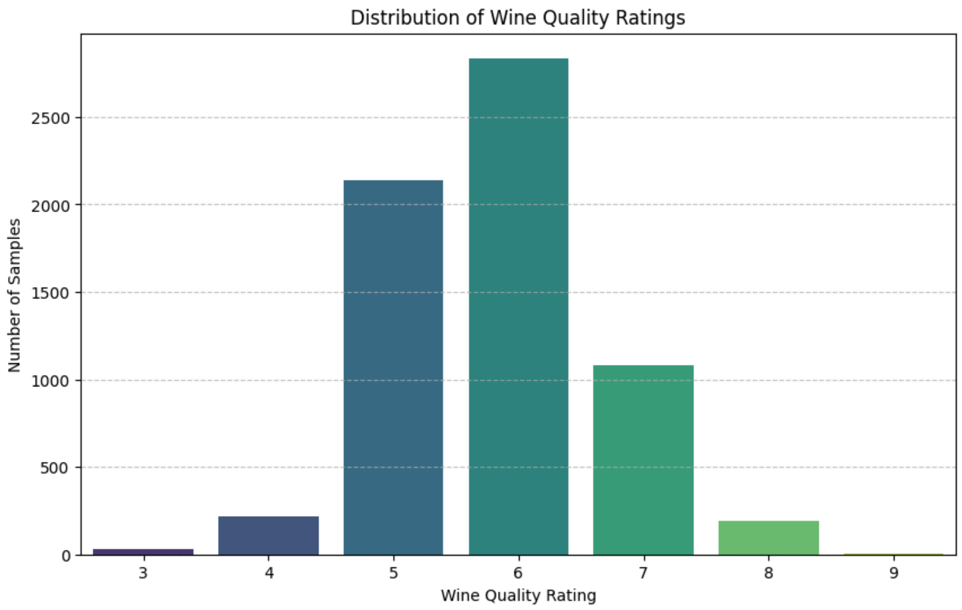
Dataset Composition:

- **Total Samples:** 6,497 (1,599 red wines and 4,898 white wines)
- **Features:** 12 physicochemical attributes + 1 categorical variable (`color`) + target quality score
- **Target:** Wine quality score (integer between 0 and 10)

Data Analysis and Visualization

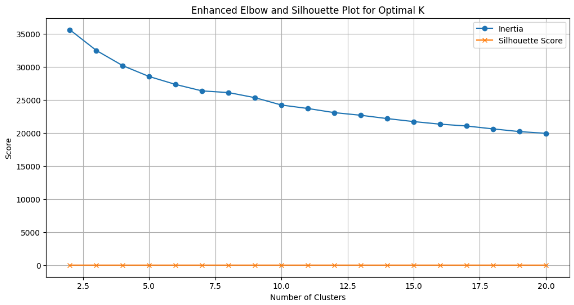
1. Distribution of Wine Quality

The first plot illustrates the distribution of wine quality ratings, ranging from 3 to 9.



Most wines are concentrated between quality ratings of 5 and 6, with 6 being the most common. This right-skewed distribution indicates that higher-quality wines (8-9) are relatively rare, while very low-quality wines (3-4) are also uncommon. This distribution provides an initial understanding of the dataset's imbalance, with the majority of wines being of moderate quality.

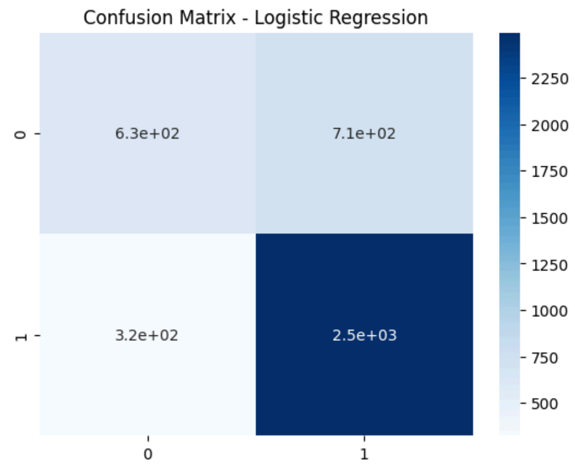
2. K-Means Clustering



The second plot presents the results of K-Means clustering. The Elbow method and Silhouette analysis indicated that three (3) clusters provided the best separation. These clusters categorize wines based on their chemical properties, including alcohol content, acidity, and residual sugar. Wines in each cluster share similar chemical characteristics, suggesting that certain chemical profiles are associated with specific quality ranges

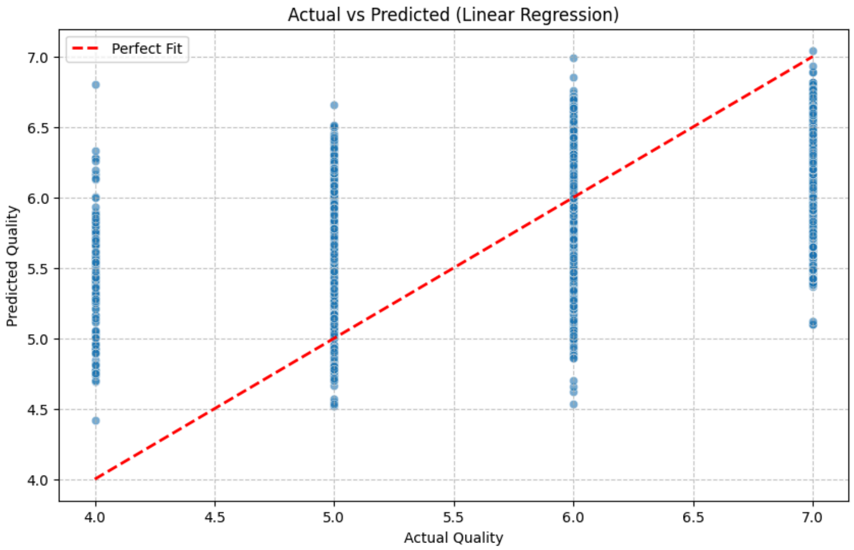
3. Logistic Regression

The third plot displays the performance of a Logistic Regression model for classifying wine quality into two categories: Good (quality > 5) and Bad (quality <= 5).



The model achieved an accuracy of 75%, with a clear Confusion Matrix showing that it effectively distinguishes higher-quality wines (precision = 78%, recall = 89%) but struggles with lower-quality wines (precision = 66%, recall = 47%). This imbalance reflects the skewed distribution of the data.

4. Linear Regression and Regularized Regression



The fourth plot compares Actual vs. Predicted wine quality using Linear Regression. The model demonstrated limited predictive power ($R^2 \approx 0.28$), indicating that wine quality is influenced by complex factors beyond the available features. Regularized regression (Ridge and Lasso) was applied to stabilize the model, but the overall predictive performance remained modest, suggesting that chemical properties alone cannot fully predict wine quality.

Conclusion

This analysis reveals that while chemical properties provide insights into wine quality, their predictive power is limited. K-Means clustering effectively groups wines based on chemical profiles, but the regression models highlight the complexity of predicting wine quality. To improve prediction accuracy, advanced techniques such as non-linear models or additional features may be required.