Nikhil Khanchandani

SEMMA Process Red Wine Quality Dataset Analysis Report

1. Sample: The dataset used is the Red Wine Quality dataset from Kaggle. It contains 1599 entries with features like acidity, sugar levels, pH, alcohol content, and a target variable quality representing the wine's quality score. We used the entire dataset to maintain the richness of the data and ensure a comprehensive analysis.

2. Explore: We conducted exploratory data analysis (EDA) to understand relationships between variables and distributions. The correlation matrix revealed significant correlations:

- Alcohol content showed a strong positive correlation with wine quality (0.48).
- Volatile acidity was negatively correlated with quality (-0.39), indicating that higher acidity often led to lower wine quality.

Visualizations, such as the alcohol content distribution and its relationship with wine quality, confirmed these trends. Higher alcohol content was associated with better wine quality.

3. Modify: We handled class imbalance using SMOTE, which oversampled the minority classes. This step improved the ability of machine learning models to predict underrepresented wine qualities. Features were standardized using standard scaling to ensure uniformity across all variables.

4. Model: We tested several models, including Logistic Regression, Decision Tree, and Random Forest:

- Logistic Regression achieved an accuracy of 39.68%, with moderate performance for class 5 and poor performance for other classes.

Nikhil Khanchandani

SEMMA Process Red Wine Quality Dataset Analysis Report

- Decision Tree improved accuracy to 49.68%, but still struggled with minority classes.
- Random Forest yielded the best accuracy of 63.43% and performed well across most classes, especially class 5 and class 6.

The balanced accuracy for Random Forest was 44.83%, indicating room for improvement in handling class imbalance, despite SMOTE. The ROC curve for multiclass classification showed reasonable AUC values for some classes, but weaker performance for lower-quality wines (Class 3 AUC = 0.29).

5. Assess: Random Forest was the best-performing model overall but struggled with precision in minority classes. The AUC score for Class 5 was 0.78, indicating good discrimination for this common quality level. However, for rarer wine qualities like 3 and 4, the model showed poor AUC scores, necessitating further improvements.

Conclusion: The SEMMA process provided a structured framework for analyzing the Red Wine Quality dataset. While Random Forest performed reasonably well, there is still scope for improvement in handling minority classes. Future work could involve hyperparameter tuning and exploring ensemble methods like Gradient Boosting for better performance.