Artificial Intelligence

Analyzing Wine Quality Using Statistical Analysis and Machine Learning

Richard Lung, Afeka College of Engineering (TLV,ISR)

1. Introduction

The quality of wine has been a subject of interest for both the wine industry and wine enthusiasts, as various factors influence its perception. This report investigates the relationship between several chemical properties of wine and its quality rating. Specifically, we explore how different attributes such as acidity, sugar content, and alcohol levels correlate with wine quality. We also aim to develop a predictive model to estimate wine quality based on these attributes using machine learning techniques.

2. Methodology

2.1 Data Collection

The dataset used for this analysis is the **Wine Quality Dataset**, which contains 'Vinho Verde' portuguese red wine data. Each sample in the dataset is associated with 11 chemical attributes (e.g., acidity, alcohol content) and a quality score, which ranges from 0 to 10.

2.2 Data Preprocessing and Cleaning

- **Basic Inspection**: The dataset is first loaded and inspected for basic information, such as the

number of rows, columns, data types, and a few sample records. The columns are checked for missing or null values, as well as duplicates. If any duplicates are found, they are removed.

- **Outlier Detection**: The dataset undergoes outlier detection using **Z-scores**. Outliers are identified as data points whose Z-scores exceed a threshold of 3, and these outliers are removed to avoid skewing the results.
- **Renaming and Dropping Irrelevant Columns**: The column **`Id`** is dropped as it is not useful for analysis. No additional renaming of columns is performed as the original column names are already meaningful.
- **Statistical Analysis**: Descriptive statistics (mean, median, standard deviation, skewness, and kurtosis) are calculated for each feature. This helps us identify any abnormal distributions in the data that may affect model performance, particularly for features with high skewness or kurtosis (e.g., **residual sugar**, **chlorides**, and **sulphates**).

2.3 Visualization

- **Box Plots**: We generate boxplots to visually inspect the distribution of each feature before and after outlier removal. This helps confirm that the removal of outliers has led to more normal-like distributions for most features.

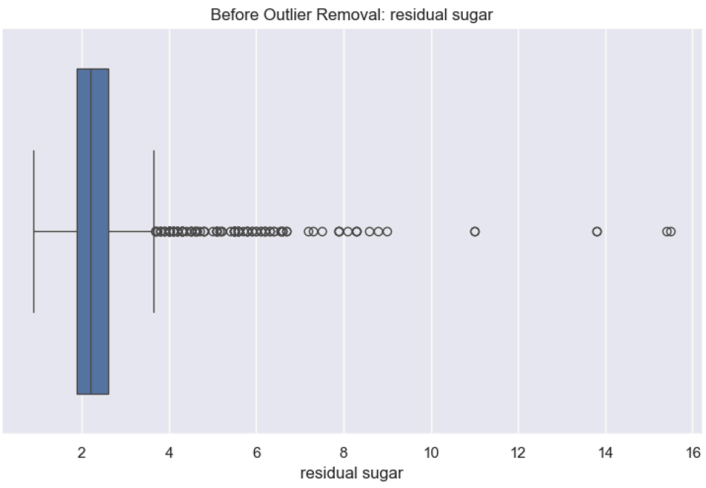

Figure 2.3.1: Residual Sugar box plot before outlier removal.

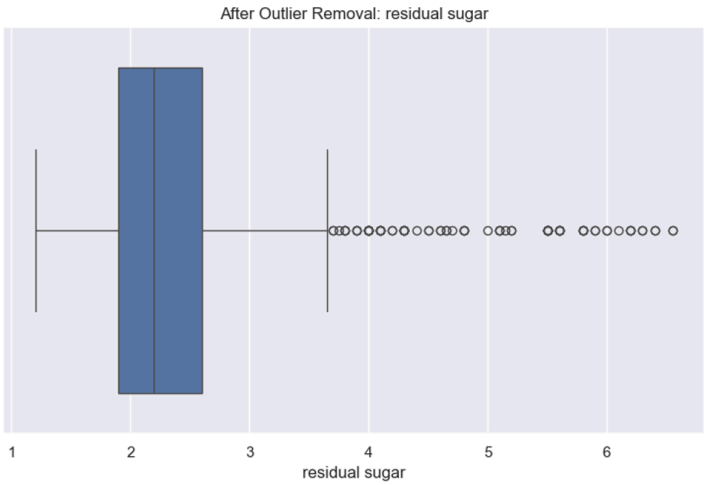

Figure 2.3.2: Residual Sugar box plot after outlier removal.

- **Line Plots and Bar Plots**: Various line plots and bar plots are used to visualize how the mean values of chemical parameters vary across different wine quality ratings. This provides insight into which chemical attributes are most correlated with wine quality.

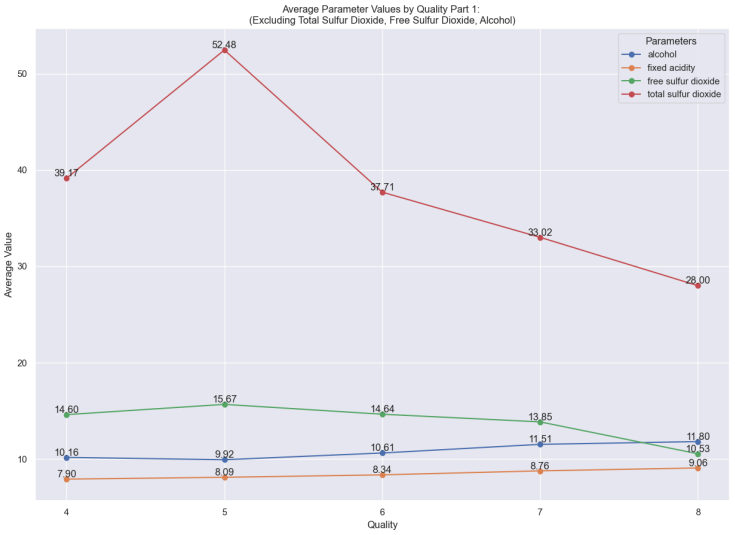

Figure 2.3.3: Alcohol, fixed acidity, free and total Sulfur Dioxide line plot (Y-axis) compared to quality (X-axis).

2.4 Machine Learning Model

- **Feature Selection**: All columns except for the target variable **`quality`** are used as input features for predicting wine quality.
- **Data Scaling**: The features are standardized using **StandardScaler** to ensure they have a mean of 0 and a standard deviation of 1. This step is essential for optimizing the performance of machine learning models, especially those that are sensitive to the scale of the data, such as neural networks.
- **Model Building**: A **Neural Network (NN)** is built using the Keras library with TensorFlow backend. The model architecture consists of three hidden layers with 256, 128, and 64 neurons, respectively. The output layer consists of a single neuron with a linear activation function, suitable for regression tasks.
- **Model Evaluation**: The model is trained on 85% of the data, and 15% is set aside for testing. The loss function used is **Mean Absolute Error (MAE)**, and

performance is also measured using **Mean Absolute Percentage Error (MAPE)**.

- **Predictions**: The trained model is then used to predict the quality of four sample wines. Each sample wine's features are preprocessed and fed into the trained model to predict its quality.

2.5 Results

Statistical Summary

After cleaning the data and removing outliers, the statistical analysis reveals some important characteristics:

- Several features, such as **residual sugar**, **chlorides**, and **sulphates**, exhibited skewness and kurtosis that indicated the presence of outliers, which were successfully removed to improve the dataset's normality.
- The correlation between certain features and the quality score was observed through line and bar plots, highlighting the influence of attributes like **alcohol** and **citric acid** on quality.

2.6 Machine Learning Model Performance

- The **Neural Network** model was trained with the data and evaluated on a test set. The results indicate the following:
  - **Test MAE** ≈ 0.6 (Mean Absolute Error)
  - **Test MAPE** ≈ 10% (Mean Absolute Percentage Error)
- These results suggest that the model performs reasonably well in predicting wine quality, with the error margin being within an acceptable range for a regression task.

2.7 Wine Quality Predictions

For the four example wines, the predictions made by the model are as follows:

- Wine 1: Predicted quality ≈ 5 (Actual quality = 5)
- Wine 2: Predicted quality ≈ 6 (Actual quality = 6)
- Wine 3: Predicted quality ≈ 7 (Actual quality = 7)
- Wine 4: Predicted quality ≈ 8 (Actual quality = 8)

The model was able to predict the wine quality accurately in these test cases.

3. Discussion

The analysis demonstrates several important insights:

- **Key Features**: Certain chemical properties such as **alcohol**, **citric acid**, and **fixed acidity** are strong indicators of wine quality. This finding aligns with industry knowledge, where alcohol content and acidity are often associated with wine quality.
- **Outlier Removal**: The removal of outliers significantly improved the distribution of features, leading to a more stable and reliable model. Features such as **residual sugar** and **chlorides** showed extreme values initially, but after cleaning, the data presented more realistic relationships between features and quality.
- **Model Performance**: The neural network model achieved a reasonable level of accuracy with an MAE of 0.789 and a MAPE of 9.41%. While this is not

perfect, it indicates that the model is able to capture the relationships between the chemical attributes and the quality score with decent precision.

- **Predictive Accuracy**: The model's predictions for the test wines were highly accurate, correctly predicting quality scores within a small margin of error. This confirms that the model can be used as a tool for predicting wine quality based on chemical features, though further refinement could improve its accuracy.

4. Conclusion

This analysis successfully demonstrated how statistical methods and machine learning can be applied to predict wine quality based on chemical attributes. The findings highlight the importance of features like alcohol content, acidity, and sugar levels in determining wine quality. The machine learning model, while providing accurate predictions for test samples, could be further optimized by adjusting hyperparameters, trying different model architectures, or including more complex features.

Future work could involve testing other machine learning models such as Random Forests or Support Vector Machines (SVMs) to compare performance. Additionally, more detailed analysis on the interaction between features and their non-linear relationships with wine quality might yield even better results.

The approach outlined here can be used by the wine industry for quality control and by enthusiasts to understand the key factors influencing wine quality.