Red Wine Quality Analysis Report

Objective

The goal of this analysis was to understand the factors influencing the quality of red wine based on various chemical attributes and to create a model that can classify wine quality effectively. This report covers data exploration, analysis, and model development, along with insights derived from each step.

---

1. Data Inspection and Setup

The dataset was initially inspected to understand the variables and their ranges, including acidity, residual sugars, chlorides, alcohol, and others. This inspection was crucial to familiarize ourselves with the data and check for any anomalies or missing values.

Insight: The data was complete, with no missing values, which allowed us to proceed directly with exploratory analysis.

---

2. Exploratory Data Analysis (EDA)

2.1 Correlation Analysis

A correlation matrix was created to examine the relationships between the chemical properties and wine quality. This analysis helped identify which attributes might have the most significant impact on quality.

Insights:

- **Alcohol content** showed a positive correlation with quality, suggesting that wines with higher alcohol content tend to be rated better.
- Attributes like **density** and **chlorides** were negatively correlated with quality, meaning that higher levels of these might adversely affect wine quality.
- Other variables showed weaker correlations, indicating that no single factor overwhelmingly determines wine quality, but multiple attributes contribute collectively.

2.2 Distribution Analysis

Histograms of each attribute were plotted to understand their distributions. This step allowed us to observe the variability and patterns within each variable.

Insights:

- Many variables exhibited normal distributions, while others, such as **volatile acidity**, were skewed. This indicated some natural variability in the dataset.

- These distributions suggested that while most wines have a typical range for attributes like acidity and sulfur dioxide, a few wines diverge, likely affecting their overall quality.

2.3 Outlier Detection

Boxplots were used to identify outliers across different variables. Outliers can sometimes indicate data entry errors or unusual conditions, but in the context of wine, they might represent unique wines with distinct qualities.

Insights:

- Certain variables, like **total sulfur dioxide** and **chlorides**, had notable outliers. While not removed, these were noted as they might influence the model's ability to generalize.

---

3. Data Preprocessing

To prepare the data for modeling, the features were standardized to ensure that each variable contributed equally to the model. The quality score was then converted into a binary target variable, classifying wines as either "high quality" or "low quality" based on a threshold.

Insights:

- Standardizing helped mitigate the impact of different scales across variables, improving the model's performance.
- Binarizing the target variable allowed us to simplify the classification task and focus on distinguishing between broadly high and low-quality wines.

---

4. Model Development

A Random Forest model was chosen to predict wine quality based on the chemical attributes. This model is well-suited for classification tasks and offers insights into feature importance.

Insights:

- **Random Forest** achieved a good balance between accuracy and interpretability, making it ideal for this dataset.
- By evaluating feature importance, **alcohol** emerged as the most critical variable in predicting quality, followed by **volatile acidity** and **sulfates**.

---

5. Model Evaluation

The model's performance was assessed using a confusion matrix and an ROC curve, which provided insights into its accuracy, sensitivity, and specificity.

Key Metrics:

- **Accuracy**: Approximately 92%, indicating that the model correctly classified wine quality in most cases.
- **AUC (Area Under Curve)**: 0.92, which reflects the model's ability to distinguish between high and low-quality wines.

Insights:

- The high accuracy and AUC score suggest that the model is effective at predicting wine quality based on the available chemical properties.
- Sensitivity and specificity metrics indicated that the model was better at identifying high-quality wines but occasionally misclassified lower-quality ones.

---

6. Feature Importance

An analysis of feature importance highlighted the top predictors of wine quality.

Insights:

- **Alcohol content** had the highest influence, supporting the hypothesis that higher alcohol levels contribute positively to perceived wine quality.
- **Sulfates** and **volatile acidity** also played significant roles, indicating that these attributes are critical for wine quality assessment.

Conclusion

This analysis demonstrated that red wine quality is influenced by a combination of chemical properties, with alcohol, sulfates, and volatile acidity being key indicators. The Random Forest model proved to be effective in classifying wine quality with high accuracy and interpretability. These findings could inform wine producers on which chemical adjustments might improve overall wine quality.

Recommendations

- Wine producers may focus on optimizing alcohol and sulfate levels to improve quality.
- Further analysis could explore nonlinear relationships or interactions between variables to refine the model and insights.