Wine-Quality-Prediction-Documentation

Table of Contents

---

1. Introduction

Welcome to the Wine-Quality-Prediction project documentation! This project focuses on predicting the quality of wines of the Vinho Verde variety based on physiochemical attributes. Wine certification involves various tests, and this model aims to predict wine quality on a scale of 1 to 10. The predictions can be valuable to certification bodies, wine producers for quality improvement, and consumers for informed choices.

2. Technologies Used

This project leverages the following technologies and libraries:

- Numpy

- Pandas

- Matplotlib

- Seaborn

-sklearn

3. Description of Qualities

The dataset includes various physiochemical attributes that are essential for wine quality prediction:

1. **Alcohol**: The alcohol content in wine.
2. **Volatile Acidity**: The acetic acid content, contributing to vinegar-like taste.
3. **Sulphates**: An additive contributing to SO2 levels, acting as an antimicrobial and antioxidant.
4. **Citric Acid**: Acts as a preservative, increasing acidity for freshness and flavor.
5. **Total Sulfur Dioxide**: The total amount of SO2.
6. **Density**: Sweeter wines tend to have higher density.
7. **Chlorides**: The amount of salt.
8. **Fixed Acidity**: Non-volatile acids that don't evaporate easily.
9. **pH**: The level of acidity.
10. **Free Sulfur Dioxide**: Prevents microbial growth and wine oxidation.
11. **Residual Sugar**: The amount remaining after fermentation stops.

Note: Features "Fixed Acidity" and "Residual Sugar" may not be informative for quality prediction.

4. Data Analysis

- **Correlation Analysis**:
  - Volatile acidity, chlorides, and pH are negatively correlated with quality, indicating that higher quality is associated with lower values of these features.
  - Free sulfur dioxide and total sulfur dioxide are highly correlated (correlation coefficient = 0.67).
  - Some features have correlations < 0.5 with quality, suggesting they may be removed, although outlier values are retained due to the need for minute-level accuracy.

5. Machine Learning Models

<u>For red wine</u>

- **Model Training Accuracy**:
  - Logistic Regression: 0.8733
  - K Nearest Neighbor: 0.8984
  - Support Vector Machine (Linear Classifier): 0.8538
  - Support Vector Machine (RBF Classifier): 0.8937

- Gaussian Naive Bayes: 0.8358

- Decision Tree Classifier: 1.0

- Random Forest Classifier: 0.9906

Conclusions:

- Outliers were not removed, and relevant features were retained to achieve high accuracy, as minute-level accuracy is essential for wine quality prediction.

- Random Forest Classifier performed the best, but other models (Logistic Regression, KNN, SVC) also had comparable scores.

- Naive Bayes exhibited the lowest accuracy, making it less suitable for wine quality prediction.

- Performance tuning using techniques like Grid Search is recommended for further improving model accuracy.

6. Conclusion

In conclusion, the Wine-Quality-Prediction project is a valuable tool for predicting the quality of wines based on physiochemical attributes. While Random Forest Classifier stands out as the best model, other models also offer competitive results. Keep in mind that there is always room for improvement in data science projects, and further enhancements, such as outlier removal and feature extraction, may yield different results. This documentation provides an overview of the project and its key findings, serving as a valuable resource for understanding and utilizing the model.