Analysing the Wine Quality Dataset with Python

Yuvraj Singh Sran

Date

10.10.23

Github profile link

https://github.com/YuvrajSinghSran/Analysing-the-Wine-Quality-Dataset-with-Python-and-M.L

Project Description

My Wine Quality Prediction project aims to harness the power of data science and machine learning to predict the quality of wines based on various attributes. Wine quality is a complex and multifaceted characteristic influenced by factors such as acidity, alcohol content, residual sugar, and more. By analyzing a dataset containing information about different wine samples, including their chemical composition and sensory properties, I am building a predictive model that can assess and classify wines as either high or low quality.

THE PROCESS

1. **Data Import:** The code imports necessary libraries and loads two datasets, one for red wine and one for white wine.
2. **Data Visualization:** It visualizes the correlation between features in both red and white wine datasets using heatmap plots.
3. **Data Splitting:** The code splits the data into training and testing sets for both red and white wines.
4. **Feature Scaling:** It standardizes the features in the datasets to ensure consistent scaling.
5. **Model Training:** The code uses Random Forest Regressor models to train on both red and white wine datasets.
6. **Model Evaluation:** It calculates Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared scores to evaluate the model's performance on the test data for both red and white wines.
7. **Feature Importance:** The code determines the importance of features for predicting wine quality and displays them in sorted order
8. **Outlier Detection:** It identifies poor-quality and excellent-quality outliers in both red and white wine datasets.
9. **Commented Outliers:** There are commented sections to print the detected outliers in the datasets.

QUESTIONS

1. Load the Wine Quality dataset into a Python environment.

I have loaded the Dataset into my model using pandas module.

2. Clean and prepare the data.

I prepared the data with the help of Standard Scaler in which I used feature scaling and did the necessary changes and as for the data cleaning part we can remove outliers such as excellent and poor wine for better results.

3. Explore the data using visualizations and statistical analyses.

I have used different libraries for this purpose such as matplotlib.pyplot and seaborn
And Visualized the data with the use of Colormap and heatmap function.

4. Answer the following questions using Python:

o What is the distribution of the wine quality scores?

I applied the Random Forest Algorithm and the features that affect the least in Red Wine is-(free sulfur dioxide: 0.0472) and the feature affecting the least in White Wine is-(citric acid: 0.0595)

o What is the relationship between the different features and wine quality?

For red Wine:

1. Most relevant feature: **alcohol->0.2709**
2. Least relevant feature: **free sulfur dioxide->0.0472**

For White Wine:

1. Most relevant feature: **alcohol->0.2391**
2. Least relevant feature: **citric acid->0.0595**

o What are the most important factors that influence the quality of wine?

For red Wine:

Alcohol->0.2709
Sulphates->0.1484
Volatile acidity->0.1115

For White Wine:

Alcohol->0.2391
Volatile acidity->0.1247
Free sulfur dioxide->0.1185

Thank You

MADE BY: YUVRAJ SINGH SRAN