Predicting Wine Quality with Linear Regression

Introduction

The goal of this analysis is to predict the quality of red wine using a Linear Regression model and explore the dataset to gain insights into the relationships between different features. This documentation outlines the steps taken to achieve this goal.

Data Analysis Libraries

- Import necessary libraries for data analysis, visualization, and modeling, including NumPy, Pandas, Scikit-Learn, Matplotlib, and Seaborn.

Data Preparation

- Load the Wine Quality dataset from a CSV file.

- Examine the dataset using `df.describe()` to get summary statistics.

- List the features in the dataset.

Data Analysis

- Display the first 5 rows of the dataset to understand its structure.

- Check the data types of each feature to identify numerical attributes.

- Use `df.describe(include="all")` to provide a summary of the dataset.

- Check for missing values in the dataset (there are none).

Exploratory Data Analysis (EDA)

Question 1: What is the distribution of the wine quality scores?

- Visualize the distribution of wine quality scores using a histogram.

Question 2: What are the relationships between the different features?

- Create a correlation matrix and display it as a heatmap to visualize feature relationships.

Question 3: Are there any outliers in the data?

- Create box plots for each feature to identify potential outliers.

Model Building

- Split the dataset into training and testing sets using Scikit-Learn's `train_test_split`.

- Create a Linear Regression model using Scikit-Learn's `LinearRegression()`.

Model Evaluation

Question 4: What is the accuracy of the linear regression model?

- Calculate and print the accuracy of the Linear Regression model.

Question 5: What are the most important features for the linear regression model?

- Display the importance of each feature in predicting wine quality.

Question 6: What is the MSE of the linear regression model?

- Calculate and print the Mean Squared Error (MSE) of the model.

- Calculate and print the Root Mean Squared Error (RMSE) of the model.

Question 7: What is the R-squared of the linear regression model?

- Calculate and print the R-squared (R2) of the model.

Model Improvement

- Suggest ways to improve the model's performance, such as feature selection, data normalization, or trying other regression algorithms.

In order to improve the performance of the linear regression model, you can:

-> Perform feature selection to include only the most relevant features.

->Apply data normalization or standardization to ensure that all features are on a similar scale.

-> Explore other regression algorithms and compare their performance.

Limitations of the Model

- Discuss the limitations of Linear Regression, including assumptions and sensitivity to outliers.

The limitations of the linear regression model include:

-> Linearity assumption: Linear regression assumes a linear relationship between the features and the target variable.

-> Sensitivity to outliers: Linear regression can be sensitive to outliers, which can affect the model's performance.

-> Independence of features: Linear regression assumes that the features are independent of each other, which may not always hold true.

-> Normality assumption: Linear regression assumes that the residuals are normally distributed.

Implications

- Explain the real-world implications of the findings, such as how this model can be used in the wine industry to predict wine quality based on its features.

Based on the findings, we can predict wine quality using the linear regression model and

identify the most important features that influence wine quality. This information can be

valuable for winemakers to understand the factors affecting wine quality and make informed decisions in the production process.