Exploratory Data Analysis Report

Soudabeh Bolouri

Introduction:

In this exercise we will have a look at the raw data before doing any machine learning. The objective is to familiarize yourself with the data, see if there are any potential problems or obstacles to applying machine learning, and if and what preprocessing may be helpful. We want to explore the distribution of selected features within the "winequality-white.csv" dataset before and after applying a normalization technique using the MinMaxScaler. We finally showed the distribution of feature values before and after preprocessing.

Dataset Description:

The dataset that we utilized in this exercise is the "Wine Quality" dataset, precisely the "white" wine version. It contains data on different features of white wines, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates, and alcohol. The dataset includes a totality of 4,898 rows and 12 features. The "quality" of the wine is the target variable, which we desire to predict. Also, we did not find any missing values in the dataset.

Experimental Setup:

We set up the experiment as follows using the Python programming language:

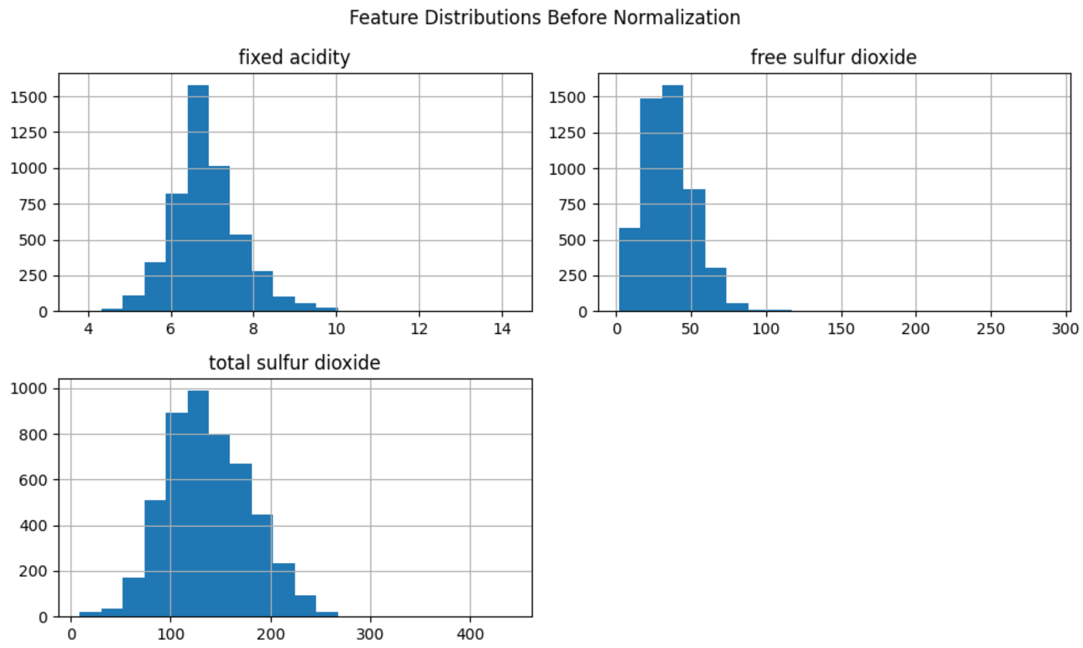
For this analysis, three features were chosen for exploration: 'fixed acidity', 'free sulfur dioxide', and 'total sulfur dioxide'. We visualized the distributions of the selected features before normalization using histograms. The histograms illustratethe spread and frequency distribution of each feature's values across the dataset.

After that a MinMaxScaler from the Scikit-learn library was employed to normalize the selected features. In normalization, the value ranges are rescaled (typically between 0 and 1), allowing a standard comparison of feature values.

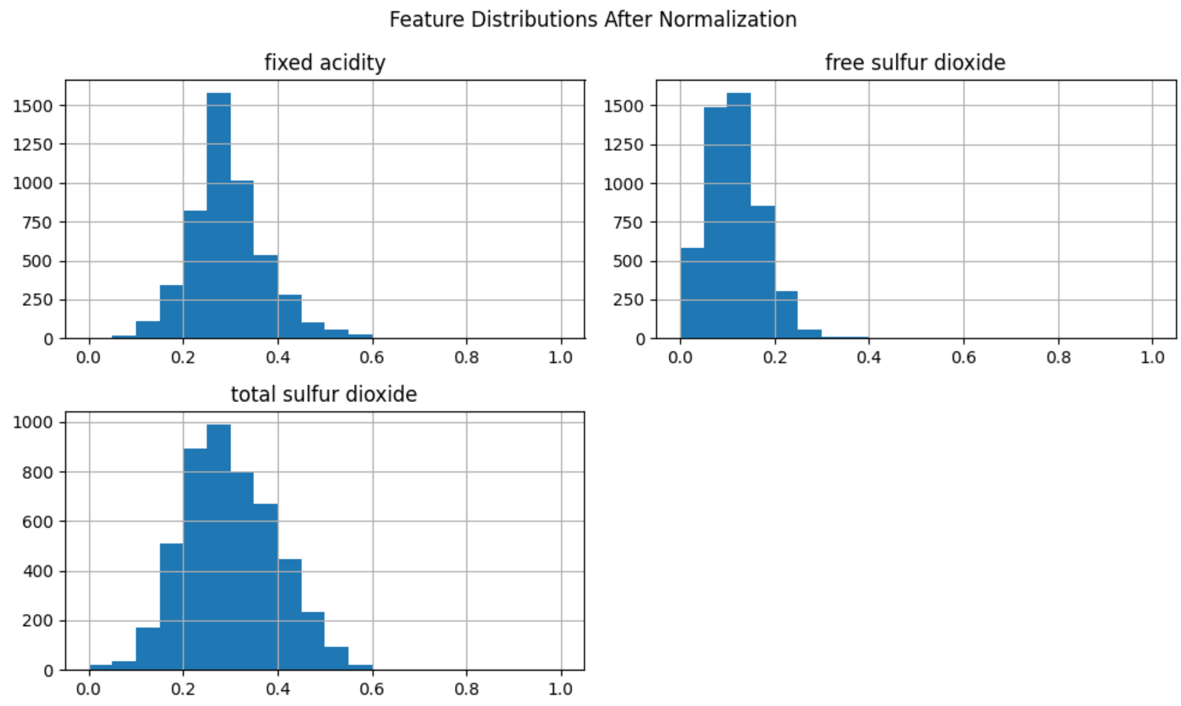
In order to explore the change in distribution following the normalization process, histograms were generated again after applying the MinMaxScaler.

Results:

When feature distributions are normalized, they can be compared on a standard scale. In the first histogram, the original distributions of the features were displayed.



After normalization and scaling, the histograms displayed a uniform distribution of features within the specified range (0 to 1).



References:

[1] https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[3] https://www.w3schools.com/python/matplotlib_histograms.asp