COSC 4557/5557 Practical Machine Learning Spring 2024
Exploratory Data Analysis
Submitted by: Iqbal khatoon

Introduction:
In this study, we set out to explore the Wine Quality dataset to better understand the dataset and potential challenges before applying machine learning techniques. Our main goal is to get familiar with the data, identify any issues that might affect our analysis, and decide if any preprocessing steps are needed to make the dataset suitable for machine learning tasks. To achieve this, we'll be using various preprocessing methods like handling missing data if any , and normalization. These techniques are aimed at resolving any issues we encounter and ensuring the dataset is in good shape for further analysis. Throughout this study, we'll discuss the challenges we face with the dataset and explain the methods we use to overcome them. Additionally, we'll evaluate the dataset both before and after preprocessing to gauge the effectiveness of our techniques. By the end of our analysis, we aim to have a clear understanding of the Wine Quality dataset and its readiness for machine learning applications.

Wine Quality Dataset:
This exercise employs the "winequality-red" dataset, containing information on different attributes of red wines. These attributes encompass fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. The dataset consists of eleven features. The target variable, denoted as "quality," assesses the wine's quality on a scale ranging from zero to ten (0-10). Based on our data analysis on the provided data set, we can ascertain that the dataset contains a total of 1599 entries, each with non-null values across all features and the label. This implies that there are no missing values present in the dataset, which is a positive aspect for our analysis. Furthermore, upon inspecting the data types assigned to each column, we observe that all features have been appropriately assigned the 'float64' data type, indicating numerical values. Similarly, the label class 'quality' comprises integer values exclusively, aligning with its assigned 'integer' data type.

Data Analysis:
Exploration of Feature Distributions Before Preprocessing:

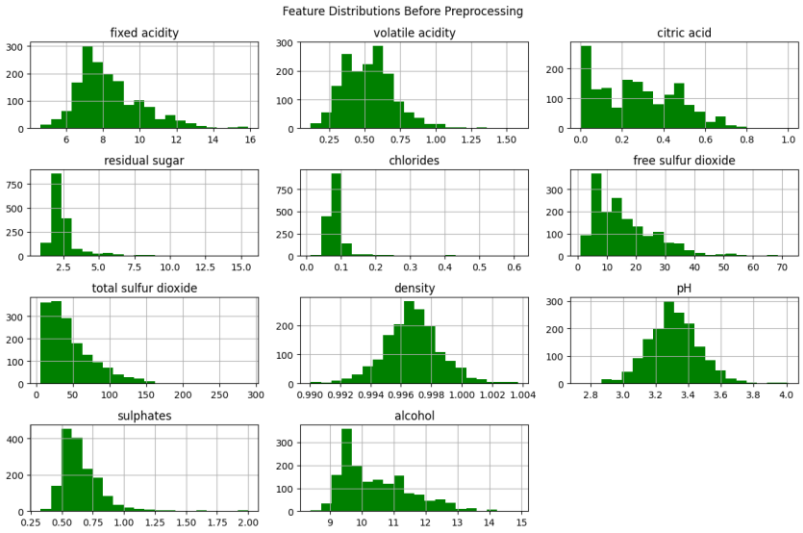

Figure 1: Feature Distributions Before Preprocessing

During our analysis, we visualized the distributions of features before preprocessing. Several noteworthy observations were made as shown in figure 1:

- **Strongly Skewed Distribution:** Residual sugar and chlorides exhibit strongly skewed distributions. This skewness indicates that the majority of values are concentrated towards one end of the distribution, potentially affecting model performance if left unaddressed.
- **Moderately Skewed Distributions:** Moderate skewness was observed in the distributions of total sulfur dioxide, sulphates, and alcohol. While not as pronounced as for residual sugar and chlorides, these features still display notable skewness that may impact model interpretation and accuracy.
- **Variations in Wine Qualities:** Different wine qualities display distinct distribution peaks for citric acid, density, sulphates, and alcohol. This suggests that these features may vary significantly across different quality grades of wine. Understanding these variations could be valuable for feature selection and model development, as they may contribute to distinguishing between wine qualities.

Skewed distributions can lead to model inefficiencies and inaccuracies, especially for models sensitive to the distribution of data. To address this issue and improve model interpretability, we employ a "logarithmic transformation" on the skewed features. This transformation helps to stabilize variance and make the data more normally distributed, which can enhance the performance of machine learning models.

Preprocessing Pipeline:

(I)Logarithmic Transformation:

Logarithmic transformation is employed to address skewed distributions within our dataset. When data exhibits a skewed distribution, meaning it is asymmetric and has a longer tail on one side, it can negatively impact the performance of certain machine learning algorithms, which assume a normal distribution. To mitigate this issue, we apply a logarithmic transformation using the log1p function. This transformation compresses large values and expands small values, effectively reducing the skewness of the distributions. The log1p function is particularly useful as it accommodates zero values by adding 1 before applying the logarithmic operation. By applying logarithmic transformation, we make the distribution of each feature more symmetrical, which can improve the performance of our models by making them less sensitive to outliers and better suited to the assumptions of certain algorithms.

(II) Standard Scaling:

Following the logarithmic transformation, we employ standard scaling to ensure consistent scaling across all features. Standard scaling transforms the data such that it has a mean of 0 and a standard deviation of 1. This step is crucial because it removes the inherent differences in the scales of different features. Features with larger scales might disproportionately influence the model, leading to biased results. By standardizing all features to have the same scale, we enable our models to interpret the importance of each feature uniformly. Standard scaling also helps algorithms converge more quickly during training, particularly for optimization algorithms that rely on gradient descent.

Results:

Figure 2 shows the after preprocessing results.

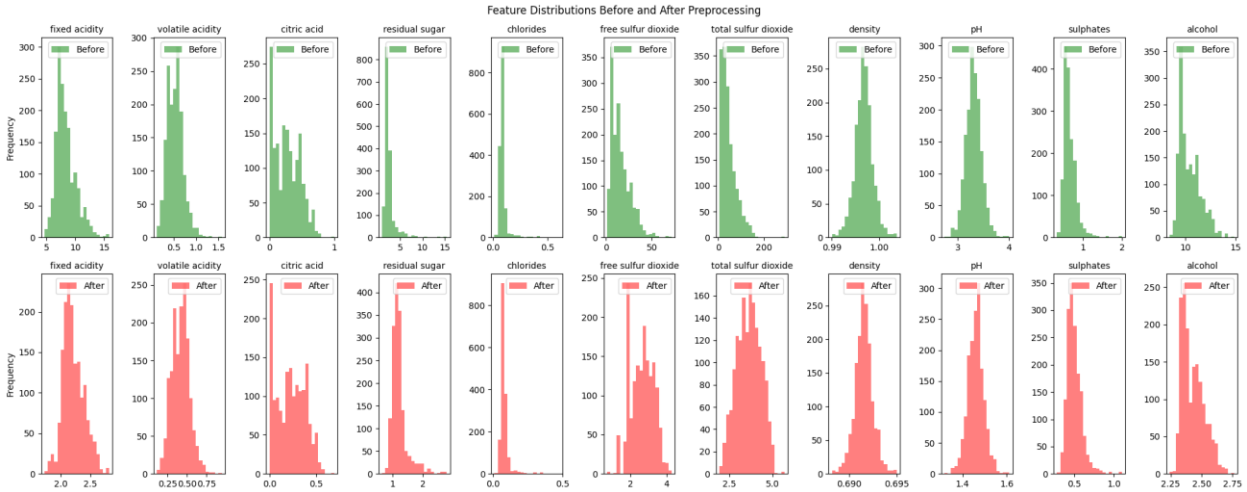

Figure 2 : Feature Distributions After Preprocessing

Conclusion:

In this study, we conducted an exploratory data analysis of the "Wine Quality" datasets before delving into the application of machine learning methodologies. Notably, to address skewed distributions observed within certain features of the "Wine Quality" dataset, we employed a tailored logarithmic transformation technique. Furthermore, standard scaling was applied to ensure uniformity across all features. By integrating these two steps into our preprocessing pipeline, we aim to improve the robustness and performance of our machine learning models, making them more capable of handling a variety of datasets and producing accurate predictions.

References:

[1] Wine Quality - UCI Machine Learning Repository

[2] **https://scikit-learn.org/stable/**