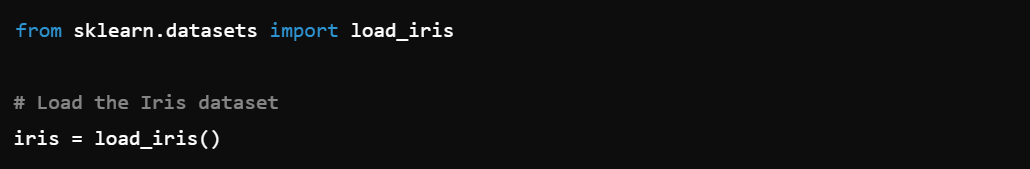
Report on Data Preparation and Data Cleaning

1. Introduction

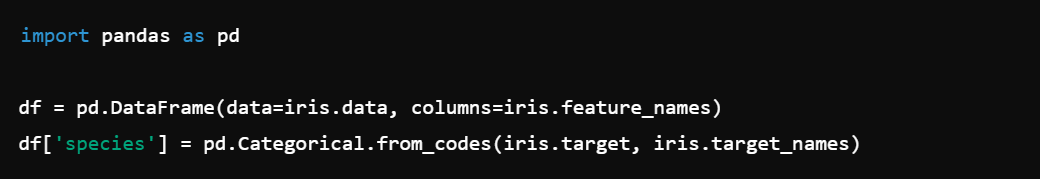
This report outlines the process of data preparation and data cleaning for the Iris dataset, a classic dataset commonly used in machine learning and statistics. The Iris dataset contains 150 samples of iris flowers from three different species, with four features measured: sepal length, sepal width, petal length, and petal width. The aim of this data preparation process is to standardize and normalize the features, ensuring that the data is in a suitable format for further analysis, such as clustering or classification.

2. Data Loading

The first step in the process is to load the Iris dataset:



This loads the dataset into a structured format with the features stored in iris.data and the target labels (species) stored in iris.target. The dataset is then converted into a pandas DataFrame for ease of manipulation:

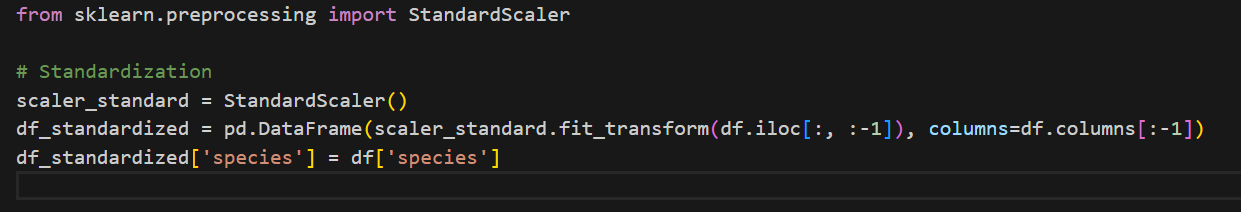


The resulting DataFrame df contains five columns: four representing the features and one for the species labels.

3. Data Standardization

Standardization is a process that transforms the data to have a mean of zero and a standard deviation of one. This is crucial when the features have different units or scales, as it ensures that each feature contributes equally to the analysis.

The code performs standardization using StandardScaler from sklearn.preprocessing:

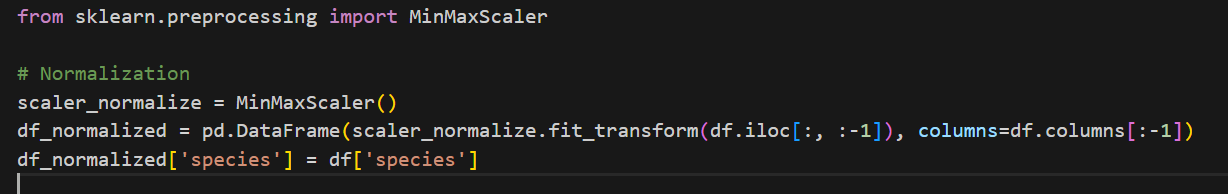


**Output**: The standardized DataFrame df\_standardized now has transformed features with a mean of 0 and standard deviation of 1. The species column remains unchanged.

4. Data Normalization

Normalization scales the data to a fixed range, typically [0, 1]. This is useful when the goal is to constrain the features to a specific range, which can improve the performance of certain machine learning algorithms.

Normalization is performed using MinMaxScaler:



Output: The normalized DataFrame df\_normalized now has features scaled to the [0, 1] range. The species column remains unchanged.

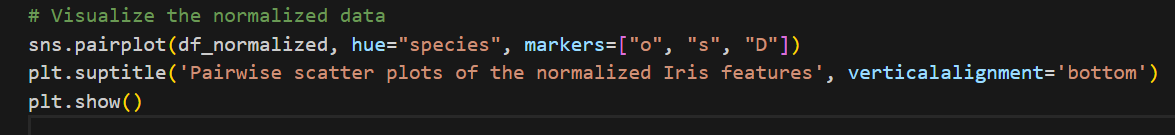
5. Data Visualization

To visually assess the effect of standardization and normalization, pairwise scatter plots are created using Seaborn's pairplot function. These plots help to understand the relationships between different features and how the data is distributed across the species after transformation.

Visualizing Standardized Data:



Visualizing Normalized Data:



Analysis:

The pairwise scatter plots for the standardized data show how the features are centered around zero, with equal variance.

The plots for the normalized data show the features scaled within the [0, 1] range, providing a different view of the data distribution.

6. Conclusion

The data preparation and cleaning process successfully standardized and normalized the Iris dataset's features, enabling consistent and fair comparisons across features. These transformations are essential steps before applying machine learning algorithms, as they can significantly impact the model's performance. The visualizations provide a clear understanding of the data's distribution after these transformations, ensuring that the data is ready for further analysis.