**Experiment 3**

**Objective:**

To perform data preprocessing and data visualization on iris dataset.

**Theory**

Data preprocessing and visualization are foundational steps in data analysis, particularly in machine learning and data science. They prepare data for deeper analysis and help reveal underlying patterns. Using the Iris dataset—a widely studied dataset in data science that includes 150 observations of iris flowers—allows us to practice these techniques effectively. This dataset contains four numerical features: sepal length, sepal width, petal length, and petal width, as well as a categorical target label representing the species of the iris flower: Iris-setosa, Iris-versicolor, and Iris-virginica.

**Data Preprocessing** involves cleaning and transforming data to improve its quality and suitability for analysis. Preprocessing can include multiple steps, such as handling missing values, scaling features, encoding categorical data, and removing duplicates. In the Iris dataset, although no missing values are present, it is still useful to check the dataset for any irregularities. Scaling or normalizing the features is also essential, especially if the data will be used in models sensitive to feature scales, such as k-nearest neighbors. Two popular scaling methods are normalization, which transforms features to a [0, 1] range, and standardization, which scales features to have a mean of 0 and a standard deviation of 1.

**Data Visualization** provides a visual interpretation of data and allows us to identify patterns, trends, and relationships within the dataset. Visualizations make it easier to compare features, observe correlations, and understand data distribution. Several plotting techniques can be applied to the Iris dataset:

1. **Scatter Plots**: By plotting feature pairs such as petal length vs. petal width and color-coding points by species, scatter plots reveal natural clusters. For example, petal length and petal width are particularly effective in distinguishing species clusters in the Iris dataset, as Iris-setosa tends to form a distinct group from Iris-versicolor and Iris-virginica.
2. **Box Plots**: Box plots show the distribution of each feature across the three iris species. They reveal the range, quartiles, and potential outliers within each feature, highlighting differences in feature distributions across species.
3. **Histograms**: Histograms help visualize the frequency distribution of each feature. For instance, a histogram of sepal length can show whether the values are normally distributed and if any values stand out as outliers.
4. **Pair Plots (or Scatterplot Matrix)**: A pair plot displays scatter plots for each pair of features, with color coding by species. This approach provides a comprehensive view of feature relationships and how they may relate to the species classification.
5. **Violin Plots**: These combine the features of box plots and histograms to show the distribution of each feature across species, providing insights into both the range and density of values within each species.

**Result**

As a result of this Experiment, we successfully wrote and executed the program to perform data preprocessing and data visualization on iris dataset.

**Learning Outcomes**

Understand and apply data preprocessing techniques and various visualization methods to clean, explore, and interpret patterns in the Iris dataset effectively.