**Experiment 2**

**Objective:**

To perform data preprocessing and data summarization on iris dataset.

**Theory**

Data preprocessing and summarization are critical steps in the data analysis pipeline, particularly when working with machine learning models. These processes ensure that data is clean, consistent, and ready for analysis. The Iris dataset, one of the most well-known datasets in data science, serves as an excellent example for demonstrating these techniques. This dataset includes 150 samples of iris flowers, each described by four features: sepal length, sepal width, petal length, and petal width. Additionally, each sample is labelled as belonging to one of three species of iris: Iris-setosa, Iris-versicolor, and Iris-virginica.

**Data Preprocessing** is a series of steps used to prepare raw data for analysis or modelling. It involves cleaning and transforming the data, handling missing values, and ensuring consistency in data formats. In the case of the Iris dataset, data preprocessing might involve verifying that each feature is numeric and consistent in scale. Since the dataset does not contain missing values, a typical first step is to check for any outliers or data inconsistencies, though the Iris dataset is known for its clean structure. However, in more complex datasets, preprocessing could include filling missing values using techniques like mean imputation, median imputation, or even more sophisticated methods like k-nearest neighbours.

Another preprocessing step is **data normalization or standardization**, especially when working with distance-based machine learning models such as k-nearest neighbours. Normalization scales features to a range, typically [0,1], while standardization scales them to have a mean of 0 and a standard deviation of 1. For the Iris dataset, where all four features are continuous and on different scales, these transformations can ensure that each feature contributes equally to model performance.

**Data Summarization** follows preprocessing and is used to understand the characteristics of the dataset. Summary statistics provide insights into the distribution and spread of the data, aiding in pattern identification. Descriptive statistics like mean, median, standard deviation, minimum, and maximum values are calculated for each feature. For instance, the mean sepal length and standard deviation help provide a quick understanding of the central tendency and spread of this feature. Summarization also includes visual techniques such as histograms, box plots, and pair plots. For example, a pair plot can illustrate relationships between sepal length, sepal width, petal length, and petal width across the different species in the Iris dataset. Box plots can reveal the distribution of each feature, highlighting any outliers or variability between species.

Furthermore, **data visualization** is a part of data summarization that provides a graphical representation of statistical summaries. In the case of the Iris dataset, scatter plots of petal length versus petal width, colored by species, can reveal clusters of species and help in visualizing decision boundaries. These visualizations are crucial when interpreting data patterns before any modeling phase.

**Result**

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

.

**Learning Outcomes**

Understand and apply data preprocessing and summarization techniques to clean, normalize, and analyse datasets, gaining insights into feature distributions and relationships.