EXPLORATORY DATA ANALYSIS ON IRIS DATASET

Short Summary Report

By:
Neha Bhivgade

Introduction:

Exploratory Data Analysis (EDA) is a crucial component in the process of analysing data, helping in comprehending the structure, patterns, and relationships in a particular dataset without using any machine learning algorithms or statistical models. Exploratory Data Analysis implies summarizing key features of a given dataset through the use of descriptive statistical methods.

In this article, an Exploratory Data Analysis is done on the Iris flower dataset, which is one of the famous datasets in Data Science. The Iris flower dataset consists of measurements taken from three species of iris flowers. By employing different data analysis and visualization tools, key information is derived related to the distribution, correlation, and classification of species. Such an EDA can help in determining key variables, thereby understanding data behaviour.

Dataset Description:

The Iris dataset consists of **150 samples** of iris flowers. Each sample contains **four numerical features**: Sepal Length, Sepal Width, Petal Length, and Petal Width, along with one categorical target variable called **Species**. The dataset includes three species: *Iris-setosa*, *Iris-versicolor*, and *Iris-virginica*.

Data Loading and Inspection:

The dataset was loaded using the Pandas library. Initial inspection using head (), shape (), and info () helped understand the structure, size, column names, and data types present in the dataset.

Data Cleaning:

The dataset was checked for missing values and duplicate records. No missing values were found, and duplicate rows (if any) were removed. Hence, the dataset did not require extensive cleaning.

Descriptive Statistics:

Statistical measures such as mean, median, minimum, maximum, and standard deviation were computed for numerical features. These statistics provided an overview of data distribution and variability among different flower measurements.

Univariate Analysis:

Univariate analysis was performed using histograms and box plots. This helped in understanding the distribution of individual features and detecting any outliers. Petal features showed more variation compared to sepal features.

Correlation Analysis:

A correlation matrix and heatmap were used to study the relationship between numerical variables. A strong positive correlation was observed between **petal length and petal width**, while sepal width showed weaker correlations.

Data Visualization:

Various visualizations such as histograms, box plots, scatter plots, violin plots, and heatmaps were used. These visual tools helped in identifying patterns, trends, and relationships within the dataset.

Key Insights:

- The dataset is clean and well-structured
- Petal features are more important than sepal features for classification
- *Iris-setosa* is easily separable from other species
- Petal length and petal width are highly correlated

Conclusion:

The EDA provided a clear understanding of the dataset and revealed important feature relationships. The insights obtained can be effectively used for further machine learning tasks such as classification and prediction.