**DevelopersHub AI/ML Engineering Interns**

**DevHub Task # 01**

**Name: Muhammad Oun Haider**

**ID: DHC-1652**

**Task 1: Exploring and Visualization of Simple Dataset**

**Abstract:**

This report documents an exploratory data analysis (EDA) of the Iris flower dataset, conducted as part of a machine learning internship. The analysis involved loading and inspecting the dataset using pandas, generating descriptive statistics, and creating visualizations including scatter plots, histograms, and box plots. Key relationships between sepal/petal dimensions across three Iris species (Iris-setosa, Iris-versicolor, and Iris-virginica) were identified through systematic data exploration using Python's data science stack (pandas, matplotlib, seaborn).

**Introduction:**

The Iris dataset is a foundational dataset in machine learning and statistics, containing measurements of 150 Iris flowers from three distinct species. This analysis serves as an introductory exercise in data exploration and visualization techniques. The dataset comprises four numerical features (sepal length, sepal width, petal length, petal width - all in centimeters) and one categorical feature (species), providing an ideal platform for practicing fundamental data inspection and visualization methods essential for any machine learning workflow.

**Objectives:**

The primary objectives were to:

* Load and inspect the dataset's structure and content
* generate summary statistics to understand feature distributions
* visualize feature relationships through scatter plots
* Examine value distributions using histograms
* Identify potential outliers through box plots.

These steps form the essential foundation for subsequent machine learning tasks by revealing data patterns, anomalies, and inter-feature relationships.

**Procedure:**

The analysis followed a structured workflow: First, the dataset was loaded using pandas' read\_csv() function. Initial inspection included checking the dataset shape (150 rows, 6 columns), column names, and first five records. Summary statistics were generated using .info() (revealing no missing values) and .describe() (showing mean, percentiles, and ranges). Visualization proceeded in three phases:

* Scatter plots (pair plot matrix and individual plots) comparing all feature combinations
* Histograms with kernel density estimates for each feature grouped by species
* Box plots showing distribution characteristics and outliers per species. All visualizations were created using matplotlib and seaborn with consistent color-coding by species.

**Discussion:**

Key insights emerged from the visualizations:

* Scatter plots revealed strong positive correlations between petal dimensions (length/width) and weaker sepal correlations. Iris-setosa showed clear separation in petal measurements.
* Histograms demonstrated bimodal distributions in sepal width, while petal measurements exhibited clear separation between species, particularly for setosa.
* Box plots identified potential outliers in sepal width (setosa and virginica) and sepal length (virginica). The most significant finding was the distinct petal characteristics of setosa versus the overlapping features of versicolor and virginica. These patterns validate the biological distinctions between species and highlight features likely to be important for classification models.

**Conclusion:**

This exploratory analysis successfully characterized the Iris dataset's structure, distributions, and feature relationships. The visualizations confirmed expected biological distinctions between species while identifying measurement variations and potential outliers. The exercise demonstrated core data science competencies: efficient data loading/inspection, statistical summarization, and informative visualization. These insights form a crucial foundation for subsequent machine learning tasks, particularly classification modeling where petal measurements emerge as potentially significant predictors. The analysis also highlighted the importance of EDA as an essential first step in any data science workflow.