Iris Data Analysis

Iris Versicolor | Iris Setosa | Iris Virginica

Dataset Overview:

- **Columns and Rows**: The Iris dataset comprises 6 columns and 150 rows, indicating a moderate-sized dataset suitable for analysis.
- **Data Types**: The "Species" column contains categorical data, representing the species of iris plants, while the remaining columns ("SepalLengthCm", "SepalWidthCm", "PetalLengthCm", and "PetalWidthCm") consist of numeric data representing various measurements of the iris flowers.
- **Non-null Entries**: All columns have non-null entries, indicating that the dataset is complete with no missing values.
- **Unique Species**: There are three unique species in the dataset: Setosa, Versicolor, and Virginica.

Methodologies Used:

- **Exploratory Data Analysis (EDA)**: EDA techniques such as pairplot, histogram, and distplot analyses were employed to explore the dataset's structure and characteristics.
- **Correlation Analysis**: Correlation heatmap was generated to quantify relationships between variables and identify potential predictors for species classification.

Pairplot Analysis:

Pairplot analysis reveals visual relationships between different variables in the dataset. The plot highlights distinct characteristics of each species.

- Species Setosa tends to have smaller sepal lengths but larger sepal widths compared to the other species.
- Versicolor species falls between Setosa and Virginica in terms of sepal length and width.
- Virginica species generally has larger sepal lengths but smaller sepal widths.
- Setosa species exhibits smaller petal lengths and widths compared to the other species.
- Versicolor species falls between Setosa and Virginica in terms of petal length and width.
- Virginica species has the largest petal lengths and widths among the three species.
- Based on the patterns observed, petal length and petal width emerge as potential features for species classification due to their distinct distributions among species.

Histogram Analysis:

- **Frequency Distribution**: Histograms provide insights into the frequency distribution of each variable in the dataset. The highest frequency of sepal length falls within a specific range, indicating a common measurement among the iris flowers. Similar observations are made for sepal width, petal length, and petal width.
- The highest frequency of sepal length falls between 5.5 and 6.
- The highest frequency of sepal width falls between 3.0 and 3.5.
- The highest frequency of petal length falls between 1 and 2.
- The highest frequency of petal width falls between 0.0 and 0.5.

Distplot Analysis:

- **Distribution Overlap**: Distplot analysis reveals the degree of overlap in the distributions of different variables. Significant overlapping is observed in sepal length and sepal width distributions, suggesting less discriminative power for species classification. In contrast, petal length and petal width distributions show relatively less overlap, indicating better potential for classification.

Correlation Heatmap:

- **Correlation Strength**: The heatmap visualizes the correlation strengths between pairs of variables in the dataset.
- Petal width and petal length exhibit high correlations.
- Petal length also shows a good correlation with sepal width.
- Petal width demonstrates good correlations with sepal length.

Conclusion:

- **Feature Selection**: Petal length and petal width are identified as promising features for species classification due to their distinct distributions and strong correlations.
- **Sepal Measurements**: Sepal measurements, while informative, exhibit significant overlap among species, potentially limiting their effectiveness in classification tasks.
- **Implications**: These findings can guide further analysis or machine learning models aimed at accurately classifying iris species based on their characteristics.