**Assignment No. 3**

**Name : Adhik Baba Sarak**

**PRN : 123B2F152**

**Problem Statement:**

**a. Find the correlation matrix on the iris dataset.**

**b. Plot the correlation plot on the dataset and visualize giving an overview of relationships among data on iris dataset.**

**Objective:** The objective of this assignment is to objective of this Python script is to compute the **correlation matrix** of the Iris dataset and visualize the relationships between its features using a **correlation plot**. Understand the linear relationships between features such as sepal length, sepal width, petal length, and petal width in the Iris dataset. Highlight which features are positively, negatively, or not correlated, which is useful in determining how features interact with each other.

**Prerequisite :**

**Python Installed**: Make sure Python is installed on your machine.

**Python Libraries**: You will need specific libraries to work with the dataset, perform calculations, and create visualizations. Install the following packages if you haven't already:

* **NumPy**: For numerical operations.
* **Pandas**: For data manipulation.
* **Seaborn**: For statistical data visualization.
* **Matplotlib**: For plotting graphs and visualizations.
* **Scikit-learn**: To load the Iris dataset.

### **Theory: Correlation and Visualization on the Iris Dataset**

#### **1. Overview of the Iris Dataset**

The Iris dataset is a widely-used dataset in machine learning and statistics, containing information on 150 flower samples from three different species of Iris: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Each sample has four features:

* Sepal Length (cm)
* Sepal Width (cm)
* Petal Length (cm)
* Petal Width (cm)

The target variable (species) indicates the class of the iris flower to which the sample belongs.

#### **2. Understanding Correlation**

Correlation is a statistical measure that expresses the strength and direction of a relationship between two variables. It provides insights into whether changes in one variable are associated with changes in another. In simple terms:

* Positive correlation: If one variable increases, the other variable tends to increase as well.
* Negative correlation: If one variable increases, the other tends to decrease.
* No correlation: No evident relationship between the two variables.

**Correlation values range between -1 and 1:**

* +1: Perfect positive correlation.
* -1: Perfect negative correlation.
* 0: No correlation (the variables are independent).

Pearson’s correlation coefficient is the most commonly used measure for linear correlation. It captures how strongly two variables are linearly related.

#### **3. Correlation Matrix**

A correlation matrix is a table that shows the correlation coefficients between many variables simultaneously. Each cell in the table represents the correlation between two variables. In the case of the Iris dataset, we will generate a matrix that reflects the correlation between the four features (sepal length, sepal width, petal length, and petal width).

#### **4. Importance of Correlation Analysis in Data Science**

* Feature Selection: Correlation analysis helps in identifying which features are important for model building. Highly correlated features can lead to multicollinearity, which can negatively impact certain models.
* Relationship Exploration: It allows the exploration of relationships between variables and helps in understanding how one variable might influence another.

For example, we may expect that petal length and petal width will be positively correlated since they both represent aspects of the petal. However, sepal width might not correlate as strongly with petal dimensions.

#### **5. Visualization of Correlation: The Correlation Plot (Heatmap)**

To visualize the correlation matrix effectively, heatmaps are used. A heatmap is a graphical representation of data where values are represented using color gradients. In a correlation heatmap:

* Darker colors represent stronger correlations (closer to -1 or 1).
* Lighter colors represent weaker correlations (closer to 0).

This visualization technique helps quickly identify the relationships between different variables in the dataset.

#### **6. Why Visualization?**

* Clarity: It’s easier to observe trends and patterns in the dataset through visual representation than by inspecting raw numbers.
* Interpretation: A heatmap highlights which features are most related, aiding in the feature engineering process, especially in machine learning.
* Anomaly Detection: You can easily detect outliers or unexpected correlations between variables.

**Algorithm:**

**Step 1: Import Required Libraries**

* Import necessary libraries like pandas for data manipulation, seaborn for visualization, and matplotlib.pyplot for displaying plots.

**Step 2: Load the Iris Dataset**

* Load the Iris dataset using load\_iris() from the sklearn.datasets module.
* Convert the dataset into a pandas DataFrame for easier manipulation.

**Step 3: Calculate the Correlation Matrix**

* Use the corr() function from pandas to compute the correlation matrix between the features in the dataset (sepal length, sepal width, petal length, and petal width).

**Step 4: Plot the Correlation Matrix**

* Use the heatmap() function from the seaborn library to create a visual plot (heatmap) of the correlation matrix.
* Set annot=True to display the correlation values in the heatmap cells.
* Choose an appropriate color map (coolwarm) to show the gradient from negative to positive correlations.

**References :**

* Iris dataset-https://www.kaggle.com/datasets/vikrishnan/iris-dataset
* Pandas Documentation
* Correlation-<https://realpython.com/numpy-scipy-pandas-correlation-python/>
* <https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e>

**Conclusion** :

In this Python script, we successfully calculated and visualized the correlation matrix for the Iris dataset, a classic dataset widely used in machine learning and statistics.By utilizing the pandas, seaborn, and **matplotlib** libraries, we were able to:

1. **Compute the Correlation Matrix**: We analyzed the relationships between the features of the Iris dataset (sepal length, sepal width, petal length, and petal width) using the corr() function. The resulting correlation matrix provided insights into how closely related each pair of features is.
2. **Visualize Relationships**: The heatmap generated through seaborn allowed for an intuitive understanding of the correlation coefficients. The visual representation highlighted strong positive correlations (e.g., between petal length and petal width) and weaker or negative correlations (e.g., between sepal width and petal length).