**NEXUS INTERNSHIP**

Data Analyst Intern - Phase 1

**Project #1: Iris Dataset Basic Analysis**

For the ‘*Exploratory Data Analysis’* part, the Python programming language has been used and implemented to carry out the EDA processes, as advised in the Project PDF provided by your organisation.

Iris Dataset is a well-known and often-used dataset in machine learning and data analysis, and it is known for its cleanliness and lack of missing values or anomalies. Therefore, in the case of the Iris dataset, explicit data cleaning was not necessary. (It is to be noted that thorough analysis of the data was still made prior.)

*Step-by-step process of EDA is as follows:*

1. *Importing the required libraries*
2. *Importing the Dataset (Iris.csv)*
3. *Basic Cleaning of the Dataset*
4. *Finding Correlation patterns and visualizing data using graphs.*
5. **Importing the required libraries:**

Few basic and the most important libraries such as

* Pandas (to read csv files)
* Seaborn (to visualize the data the analysis that has been made with the data)
* Io (to import files into the Colab environment from Local)

1. **Importing the Dataset:**

The provided dataset (Iris.csv) has been used to perform the basic analysis advised and EDA was carried out on the same.

1. **Basic Cleaning of the Dataset:**

Few basic cleaning operations such as checking and removing null values (no null values were seem to be present in the data); also, the ‘Id’ column was removed from the dataset using (.drop) operation since, the column was found to be redundant to the analysis.

In addition to this, all the entries in the ‘Species’ column which had the prefix ‘Iris-’ was replaced by removing the prefix.

1. **Finding Correlation patterns and visualizing data using graphs:**

Using the seaborn library, the correlation heat-map and the pair-plot were used to visualize the relationship between the various features of the dataset given.

For the *‘Data Visualization’* part, I have employed the use of Power BI and have carried out the visualization task with the same.

The ‘**Iris-Dataset-Visualization.pbix**’ contains two pages of visualizations that could be extracted from the Iris Dataset. Since, the features with which the visualizations could be carried out was so less; there was not much scope of using various visualization tool.

On the First Sheet, I have used the ‘Slicer’ tool with the combination of ‘Donut Chart’; in the idea, so that whenever the features (Sepal Length, Sepal Width, Petal Length, Petal Width) on the slicer tool is being altered, the Donut Chart updates it accordingly, emphasizing what species does the current input invokes to.

On the Second Sheet, I have used the Scatter Plot to plot all the points of various features that I have mentioned above. Whilst plotting the scatter plot, it was clearly seen that the species ‘Iris-Setosa’ was always standing distinct from the other species.

But, whereas the species ‘Iris-Virginica’ and ‘Iris-Versicolor’ had always been overlapping each other with a few outliers being found sparsely. So, the process of analysing between these two species could really be a hard nut to crack.

~ END OF REPORT ~