**FOUNDATIONS OF DATA SCIENCE**

**LAB ASSIGNMENT - 3**

**LINEAR REGRESSION**

This lab will use the same dataset as that of Lab 2. More information about it is given below:

Download the Iris dataset from the given link - <https://www.kaggle.com/uciml/iris/version/2>

The Iris dataset was used in R.A. Fisher's classic 1936 paper, [The Use of Multiple Measurements in Taxonomic Problems](http://rcs.chemometrics.ru/Tutorials/classification/Fisher.pdf), and can also be found on the [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/).

It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

* Id
* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
* PetalWidthCm
* Species

After downloading the .csv file, load its contents into a dataframe on a Python notebook and perform preprocessing techniques. Now, perform the following EDA operations:

1. Display a scatter plot showing the relationship between sepal length and sepal width.
2. Find how many flowers of each type of species are present in the dataframe.
3. Show how the length and width of the petals are distributed by means of a histogram.

Now, we will use the data to apply regression techniques to predict certain attributes of the Iris flowers dataset.

**Problem Statement 1** : Perform linear regression to check if the petal lengths and petal widths have a linear relationship. In order to do so, split the dataset into training and testing dataset (The testing dataset must be smaller than the training dataset – do a 75/25 split). Make sure you use the reshape function since the .fit method is expecting the x and y variables to be 2 dimensional arrays. Print coefficient, intercept, and R squared values of test and training datasets. Evaluate the model’s performance based on these values and comment on the linear relationship between the variables. Confirm your findings by plotting scatterplot with regression line.

**Problem Statement 2 :** Perform linear regression to predict the sepal length of the iris flowers and print Mean Absolute Error, Mean Squared Error, Mean Root Squared Error. Compare your predicted values with the original values to see how well your model did.