**FOUNDATIONS OF DATA SCIENCE**

**LAB ASSIGNMENT - 5**

**Dataset:**

We will be using a dataset which contains information about whether a sonar signal bounced off a rock or a metal based on a given set of conditions. The dataset has already been cleaned and preprocessed according to how we require it.

The dataset has 60 features in which the sonar signals are labeled to have bounced off of ‘M’ for metals and ‘R’ for rock. The ‘is\_M’ column has value 1 for metal and 0 for rock.

**Perform the following operations:**

1. Load the dataset into a dataframe. If it contains an index column in the start called ‘Unnamed : 0’ column at the start of the dataframe, drop the column.
2. Perform EDA to check whether the ‘is\_M’ and ‘Label’ columns match appropriately. Print the top 5 values (or head) in this comparison.
3. Split the dataset into a 70:30 train-test ratio.
4. The task is to predict the values of ‘is\_M’ based on the 60 features in the dataset.
   1. Use Logistic Regression to predict the values of ‘is\_M’. Compare the predicted values with the actual values and find its accuracy.
   2. Use kNN to predict the values of ‘is\_M’. Compare the predicted values with the actual values and find its accuracy. Loop this for values k = 1 to 10. Plot a graph to show which value of k has the highest accuracy.
   3. Use Decision Tree (Entropy or Gini coefficient) to predict the values of ‘is\_M’. Compare the predicted values with the actual values and find its accuracy.
5. Which model has the highest accuracy?