**Assignment 2**

**Develop UnSupervised Machine Learning Models**

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features.

In previous assignment, we developed supervised learning model to predict attacks on the CAN bus of a vehicle. But, in reality data used on these models are generally unlabeled (meaning, you will not be provided label column stating Attack =1 or Attack =0). Let’s develop unsupervised learning models using the same datasets.

In this assignment you are tasked to study provided files i.e bus logs, extract relevant information, explore data and come up with 2 clustering algorithms i.e., K-Mean Clustering and Isolation Forest Clustering.

The tasks are:

**Task 1: Data preparation**

You are provided three files, i.e., text files containing bus logs of 3 different scenarios

1. CAN Bus log - injection of FFF as the speed reading
2. CAN Bus log - injection of RPM readings
3. CAN bus log - no injection of messages Update your Python module to read the datasets were provided in assignment 1.

Unlike previous assignment, you will not use label data. You need to drop ‘Attack’ column from dataframe. Report must have screenshots of the data frames output (containing columns RPM & Speed) for all the 3 cases.

**Task 2 - k-mean Clustering**

K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms. A cluster refers to a collection of data points aggregated together because of certain similarities. You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. For this assignment you need to set k = 2 as we have 2 outcomes (i.e., Attack = 0 and Attack = 1)

You need to fit k-mean clustering separately for 6 datasets (1 Speed and 1 RPM dataframe for each of 3 scenarios).

1. Use scatter plot to show the fitted k-mean clusters.
2. Provide the centroids of each clusters and the accuracy of the learning.
3. Compare and discuss the scatter plots of the three scenarios from the speed datasets.

Do the same for the RPM datasets.

Report must have screenshots of 6 plots, centroids, accuracy and comparison/observations on the same. Write one or two paragraphs to report your observations.

**Task 3- Isolation Forest Algorithm**

You need to fit Isolation Forest separately for the 6 datasets (1 Speed and 1 RPM data frame for each of the three scenarios).

1. Use scatter plot to show the fitted clusters
2. Note down data point’s values (for example, RPM = 2700) detected as anomalies by IF and the accuracy of the learning.
3. Compare and discuss the scatter plots of the three scenarios from the speed datasets.

Do the same for the RPM datasets.

Report must have screenshots of 6 plots, anomalies, accuracy and comparison/observations on the same. Write one or two paragraphs to report your observations.

**Task 4 - Hidden Markov Models:** Build Hidden Markov Model to predict Attack vs no Attack.

1. Use scatter plot to show the fitted clusters
2. Note down the accuracy of the learning.
3. Compare and discuss the scatter plots of the three scenarios from the speed datasets.

Do the same for the RPM datasets.

Report must have screenshots of 6 plots, accuracy and comparison/observations on the same. Write one or two paragraphs to report your observations.

**Task 5 - Discussion** Write 1-2 paragraphs reflecting on what you have learnt from this assignment. Questions that you may consider are:

1. What were the shortcomings/limitations of using
2. unsupervised ML on this problem?
3. What could be done to make the models perform better?
4. What do you suggest doing to solve the problem?

Report your reflection in one or two paragraphs containing side-headings ’Limitations’, ’Improvements’ and ’Alternative Solutions/Suggestions’.