**ML Assignment# 1 – Data scrubbing and preparation**

**Introduction:**

The process of dealing with missing, incorrect, incomplete, insignificant, improperly formatted, or duplicated data in a dataset is called data cleaning or scrubbing. Machine learning (ML) algorithms provide a better result when the dataset used by the algorithm is well formatted and error-free. So, before deciding or applying an ML algorithm, Data scientist or Data Analyst perform following two steps:

1. Data Exploration - Understand the source data in detail and logically associate the attributes
2. Data Scrubbing – Correct errors found in the source data and format the data for the best performance of the selected ML algorithm

**Assignment and Deliverable** (Total 40 points)

This assignment is divided into two parts.

1. Part 1, the students use Assignment#1\_Part1\_Motor\_Insurance\_Fraud.csv dataset
2. Part 2, the students use Assignment#1\_Part1\_Online\_Activity.csv dataset

Perform the steps described below using Python. You can submit your Jupyter Notebook to canvas or just share the Google Colb notebook link. If there is a question to be answered in the steps below, use “add text” in the notebook and answer them right below after running your code.

**Part 1 – Data Exploration** (20 points)

Step 1: Load the Assignment#1\_Part1\_Motor\_Insurance\_Fraud.csv data into Google Colab. Read that into a DataFrame.

Step 2: Explore the dataset

* Display the dataframe
* What are all the columns? What are their data types?
* Pick any column and show it.
* Create new column “revenue” which is sum of “Num claims” and “Claim Amount Received”. Check to see if new column is there or not.
* Select any row and display that.

Step 3: Identify missing attribute field(s). Which have missing values and how many? Propose a way to resolve these missing values for those attributes.

Step4: Consider the attribute “Insurance Type”. Do you find it odd? How would you remove that attribute?

Step 5: Explore how the attributes vary or relate to each other. Calculate and visualize correlations using correlation matrix

Step6: Try to keep some and eliminate some attributes based on correlation matrix.

Step 7: Save your notebook and share it on canvas. Using text comments, write briefly your observations and inferences.

**Part 2 Data Scrubbing** (20 points)

Step 1: Correct Missing values:

1. Load the Assignment#1\_Part1\_Online\_Activity.csv data into Google Colab
2. Describe the dataset and provide
   * number of rows/columns,
   * data type,
   * attribute columns
3. Replace the attribute “Online\_Gaming“ missing value as “N”

Step 2: Data Reduction:

Remove only those rows with missing values in online\_shopping

Step 3: Replace Invalid Data:

Replace 99 with 'N' in 'Twitter'

Step 4: Attribute reduction - Reduce the number of attributes by eliminating unused attributes

Remove “Other\_Social\_Network”

Step 4: Save your notebook and upload to canvas.

Step 5: Using text cells, write briefly your observations and inferences.