Term Project

Data Science and Machine Learning in Canada

INSTRUCTOR: *Mohammad Saiful Islam*

CLASS: *AML 1104*

- *Group task*
- *Time allowed: 3 weeks*
- *Due date:*
- *Total marks: 30*

Instructions:

Follow the instructions below for exploratory data analysis and submit a report with the result and code by the due date.

- ❖ Conduct an exploratory data analysis of UCI Machine Learning data set. Use supervised and unsupervised methods.
  - ▪ Example UCI Machine Learning data sets:

    i) Heart Disease Data Set [https://archive.ics.uci.edu/ml/datasets/Heart+Disease]

    ii) Adult Data Set [https://archive.ics.uci.edu/ml/datasets/adult]

The expectations include the following from your experience in the course:

- ▪ Use Python Jupyter Notebook for the analysis.

- ▪ Data preprocessing:

  a) Load the data set into a data frame

  b) Are there any missing values in the dataset? How to handle that.

  c) Use feature selection and pruning techniques.

  d) Try to find out the existence of outlier in data; clean it if exists.

  e) Perform normalization of the selected features.

- Data visualization:
  - Use Data visualization techniques to plot in graphs.
  - Use a measure of central tendency for each feature.
  - Show the dispersion (standard deviation and IQR) of features.

- Supervised learning:
  - Explore random split of data as test and training set using Python.
  - Use the training data set to train the classification model; Binary classification should be fine for simplicity.
  - Observe the performance of the model with test data set.
  - Create a confusion matrix to present the result.

- Unsupervised learning:
  - Use K-means algorithm to find out cluster from the data set.
  - Try different number of clusters to compare the results.

❖ Deliverables

a) Completed Jupytar notebook

b) A .pdf report describing your observations.