Project Specification on Machine Learning

1. **Introduction**

   Students of CSE 4108 course have to do project work on Machine Learning. Students have to develop dataset on a particular topic and apply machine learning (ML) models and do comparative study. For the project work, students can form groups. Maximum number of members in a group can be three (03). Meaning, a group can have 01 member / 02 members / 03 members. The topic on which each group will work must be a classification topic or regression topic.

2. **What you need to submit**

   Overall, a group has to submit a .zip file of the folder that will contain
   - Dataset
   - Documentation of the Dataset
   - Code (in .py or, .ipynb formats)
   - Report

3. **Dataset**
   a. **Size of the dataset**
      i. The dataset must have at least 300 samples.
      ii. The samples in the dataset must have at least 07 features.
   b. **Documentation of Dataset**

      Each group has to write and submit a documentation about the dataset. It will include the descriptions about
      i. The size (i.e., number of samples and number of features) of the dataset.
      ii. Description of features with their units.
      iii. Citation of the websites from where data has been collected.
      iv. In case of classification dataset, distribution of the labels.
   c. **Example of datasets**

      You can have a look at the https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names for classification and Boston Dataset (toronto.edu) for regression dataset to find out how they constructed the datasets.

4. **ML Models**

   Students have to write the codes in Python for classification or, regression models. Students are permitted to use scikit-learn. At least five models have to used to comparative study.

5. **Performance metric**

   For each model, get at least four performance metric scores and do comparative study mentioning them in the report.

6. **Report**

   The report should contain
   - The description of the problem the group has worked on.

- A brief description of the dataset.
- Description of the used ML models.
- Comparison of the performance scores of the models (Use tables / plots for comparison of performance scores).
- Discussion (The conclusions the group has come to learn after model analysis).
- In case of group with more than one member, percentage of contribution of each member indicating on which each member has worked on.