Project Specification on Machine Learning

1. **Introduction**

   Each student of CSE 4108 course has to do project work individually on Machine Learning. Each student has to develop a dataset on a particular topic and apply machine learning (ML) models and do comparative study. The topic on which a student will work can be classification or regression or clustering.

2. **What needs to be submitted**

   Overall, a student 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 450 samples.
      ii. The samples in the dataset must have at least 07 features (eight attributes)
   b. **Documentation of Dataset**

      Each student has to write and submit 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**

      Each student can have a look at the IRIS Dataset [LINK] and the Boston Housing Dataset [LINK] to see how datasets are generally constructed.

4. **ML Models**

   Each student has to write the codes in Python. Students are permitted to use scikit-learn. At least five models have to be used for comparative study. At least two models must be of ensemble learning type.

5. **Performance metric**

   For each model, at least four performance metric scores have to be presented and comparative study has to be done mentioning them in the report.

6. **Report**

   The report must contain
   - The description of the problem the student 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 student has come to learn after model analysis).