

At last, we are near the end of this unit. I hope you learned something from it.

**This Assignment consists of 3 parts.**

**Part 1: Programming – 20 Marks (10 Marks for each)**

**Part 2: Report (1000 – 1500 words: 1000 if 2 members & 1500 if 3) – 10 Marks**

**Part 3: Recorded Presentation (2 mins or 3 mins) (1 min/person) – 10 Marks**

**Part 1 (Programming) (2 tasks)**

Part 1 consists of 2 tasks. You are required to implement both Task 1 and Task 2 using Python, and submit **a single file** with all the following implementation steps **for each task** (.py or. ipynb):

- 1) Data Preprocessing: Load the provided dataset and perform any necessary preprocessing steps such as normalization or standardization.
- 2) Data splitting: split datasets into 3 disjoint datasets respectively for training, evaluating, and testing.
- 3) Model Implementation: Implement a model using any existing library (e.g., sklearn, scikit-learn, pytorch, tensorflow, etc.) or designing models by yourself.
- 4) Model Training: Train the model using the training dataset.
- 5) Model Tuning, Implement hyperparameter tuning on the evaluating dataset using techniques like grid search or random search to improve the model's performance.
- 6) Model Evaluation: Evaluate the trained model on the testing dataset using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score, etc.

### **Task 1:**

Develop clustering algorithms on the 'kaggle\_Interests\_group' dataset

<https://www.kaggle.com/code/skanderhaddar/clustering-groups-of-hobbies/input>, which contains 4 groups of 6340 people and 217 hobby's and interest questions. You can choose **any two** of the following clustering methods:

- 1) K-means
- 2) Hierarchical clustering
- 3) DBSCAN
- 4) BIRCH

### **Task 2:**

Develop classifier algorithms on the 'emails' dataset

<https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv/data>, that contains related word information of 5172 randomly picked email files and their respective labels for spam or not-spam classification. You can choose **any two** of the following classification algorithms::

- 1) Logistic regression
- 2) Support Vector Machine
- 3) MLP: Multi-layer Perceptron

### **Part 2 (Report) (1000 – 1500 Words) (500/person)**

Create a concise report that details the important aspects that are related to the tasks.

(Example: List of questions you created for task 2).

**Part 3 (3 – 4 min Presentation) (1 min/person)**

Please provide a recorded presentation briefly explaining the tasks you have done.

Everyone in your group should speak and show their face while recording the presentation.

**Group submission:**

Please Zip all the files together and upload it to learnline by group. Provide student names and ID in the report. Use your group number to name the Zip file.