HIT391 - Assignment 3 (Group Assignment) **Sem 1, 2025**

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 – 2000 words: 1000 for 2 members & 1500 for 3 & 2000 for 4) – 10 Marks

Part 3: Recorded Presentation (2 mins or 4 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 spliting: Divide the dataset into three **disjoint subsets** for **training, evaluation, and testing**. For the task 2, you will need to split the provided '*train.csv'* file into two separate sets: one for training and one for evaluation.

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 cross-validation, 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.

- Objective: design a clustering method to cluster these people based on their hobbies.

- 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 for Kaggle competition 'Titanic - Machine Learning from Disaster' https://www.kaggle.com/competitions/titanic/overview.

- Objective: use machine learning algorithms to create a model that predicts which passengers survived the Titanic shipwreck, by using passenger data (ie name, age, gender, socio-economic class, etc)

- You can choose at least any **two** classfication algorithms (at least **one** is chosen from the list below):

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

Part 2 (Report) (1000 – 2000 Words) (500/person)

Create a brief report highlighting the key aspects of each task. (Example: list the problems you solved or new findings for each task)

Part 3 (2 – 5 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.

Submission Requirements (submitted by one group memeber):

Each group must upload the following sepearte files to Learnline,

1. **Two code files** named task_1.py and task_2.py
2. **A report** named assignment_3.doc or assignment_3.pdf
3. **A recorded presentation video**

Note: Please include the **names** and **student IDs** of all group members **in the report.**