In the name of god

Your final project contains sections parts:

- ➢ Classification
- ➢ Clustering
- ➢ Paper summarization

Each section includes several tasks.

Classification

The problem we're trying to solve here is to classify Letter images of handwritten digits into their 26 categories (A through Z). We'll use the UCI dataset, a classic in the machine-learning community, which has been around almost as long as the field itself and has been intensively studied.

Dataset:

Download data from: https://archive.ics.uci.edu/ml/datasets/Letter+Recognition

- Download the Data Set Description and read it carefully. This dataset contains 17 attributes, where the first column represents the label of each sample. Download the complete dataset
  letter-recognition.data

Python part

In this part, you should write python codes to classify the dataset. Follow the steps:

1. Read the dataset
2. Consider 5-fold cross-validation to determine train and test sets.
3. Apply different classifiers.
4. Classify the label of test sets.
5. Compute accuracy, recall, precision of different classifiers

Details of step 4, you should use the following classifiers:

❖ KNN:
   - Write KNN code and don't use python functions
   - Consider K=1, K=3
❖ Naïve Bayes
❖ Dtree C4.5
   - Adjust the Decision Tree Classifier parameters
   - Please plot one of the trees in fold1
   - Describe your setting parameters

- Dtree is a binary classifier and our problem is a multiclass problem. Describe how you use python codes and how you implement Dtree.

*you can use python functions for all parts except the KNN

*: Please send your homework in the form of a report. Your report must contain the description of the dataset and table for the dataset information. In addition, you should mention that either you normalize your data or not. Compare the results of classifiers and considered a table to compare and write a conclusion for your results. Describe your Python codes. Attach your Python code.

Clustering

Python part

1. Write a python program with the following steps:
    - ✓ Read a dataset
    - ✓ Apply clustering algorithm
    - ✓ Calculate the accuracy
    - ✓ Plot the clusters

2. Please select **two UCI datasets** for the clustering task that contains cluster ground truth for validation.

    https://archive.ics.uci.edu/ml/datasets.php

3. Change your python code for the following clustering algorithms:
    - ❖ K-means
        - ▪ Considered number of clusters more than the data real clusters
        - ▪ Considered number of clusters less than the data real clusters
        - ▪ Considered number of clusters equal to the real data clusters

    - ❖ Hierarchical
        - ▪ Single link
        - ▪ Complete link

*: Please send your homework in the form of a report and attach your Python code

Paper

1. Read the following paper:

Zebari, Rizgar, et al. "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction." *Journal of Applied Science and Technology Trends* 1.2 (2020): 56-70

2. Summarize the algorithms in the paper and write what you understand

Good Luck