Winter 2025 CIS*6180/Data*6300 Analysis of Big Data
Assignment #1
Total Marks: 15; Due date: February 14, 2025 (11.59 pm)

What you need to submit:
A single zip file that contains the following:

1. A PDF file containing the following:
    a. Two 3D visualization plots (one from K-Means and one from DBSCAN).
    b. Answers to the questions asked in Step 4.
2. A CSV file with three columns: Datapoint labels, K-Means labels, DBSCAN (run#2) labels
3. Your code

Programming language: Use as per the course outline.

The objective of this assignment is to learn the following: *Load and process high-dimensional data, apply K-Means and DBSCAN clustering, use PCA for dimensionality reduction, visualize results in 3D space, and analyze how parameters affect clustering.*

DataSet: You will use the provided dataset ("dp_with_labels.csv"). It contains 1550 data points, each represented as a feature vector of 4096 features. The first column contains datapoint labels (DP0001, DP0002, ..., DP1550). The remaining 4096 columns are numerical features. So, each row contains a data point label (column 1) and its respective feature vector (columns 2 - 4097).

Step 1 (2 marks): Load and preprocess the dataset. You might need to separate datapoint labels (strings) from their respective feature vectors (numerical).

Step 2 (5 marks): Perform clustering using the K-Means algorithm:

a. Apply the K-Means algorithm to the dataset, setting the number of clusters k = 4. Initialize the algorithm with multiple restarts (at least 5 replicates) to ensure convergence to optimal centroids. Use 500 as the number of iterations and 'Euclidean distance' as a distance metric for cluster assignment.
b. After clustering, apply Principal Component Analysis (PCA) to the dataset and extract the three most significant principal components (*features with the highest variance*). Use these three PCA-reduced features to generate a 3D visualization of the clustered data. Plot the clustered data points with distinct colours for each cluster and mark centroids distinctly in the 3D visualization.
    Hint: Compute cluster centroids in the 3D PCA-reduced space by taking the mean of the PCA-transformed feature vectors along the first dimension for each cluster.

Step 3 (5 marks): Perform clustering using the DBSCAN algorithm:

a. Apply the DBSCAN algorithm to the dataset. Perform two separate runs of DBSCAN with the following parameter settings:
    - Run#1: Epsilon (ε) = 2, Minimum Points (*minPts*) = 4
    - Run#2: Epsilon (ε) = 2, Minimum Points (*minPts*) = 6

b. For each run of DBSCAN, once clustering is complete, apply PCA to the dataset and use the three most significant PCA-reduced features to generate a 3D visualization of the clustered data. Plot the clustered data points with distinct colours for each cluster and mark noise points distinctly in the 3D visualization.

Step 4 (3 marks): Comparative Analysis and Discussion:

a. How does the number of clusters identified by DBSCAN vary between the two parameter settings (run#1 and run#2)? Why does increasing *minPts* affect the number of detected clusters? (*Include your answer in the pdf*)?
b. How many data points belong to each cluster for K-Means (K = 4) and DBSCAN (run#2)? - *Include your answer in the pdf*. Which algorithm appears to provide better cluster separation for this dataset? Explain why.
c. Prepare a CSV file with three columns: original Datapoint labels, K-Means labels, and DBSCAN (run#2) labels. *Example labels for K-Means and DBSCAN – Cluster1, Cluster2, etc.*