Machine Learning – Term 242
Assignment 3: Clustering

1. Dataset

As the owner of a supermarket mall, you have collected customer information through membership cards. The dataset includes basic demographic and financial attributes such as Customer ID, Gender, Age, and Annual Income. Your objective is to segment customers into meaningful groups to assist the marketing team in designing more targeted and effective strategies.

The dataset can be accessed at the following link:

https://www.kaggle.com/datasets/dev0914sharma/customer-clustering/data

2. Tasks

The objective of this assignment is to apply clustering techniques to analyze supermarket mall customers and uncover patterns in customer demographics and financial behavior using the K-means algorithm.

2.1 Centroid Initialization (10 points): Implement the K-means algorithm using K-means++ to initialize centroids for improved convergence and accuracy.

2.2 Optimal Number of Clusters (25 points): Determine the optimal number of clusters using two complementary methods:

a) Elbow Method (WCSS – Within-Cluster Sum of Squares)
- Compute WCSS for a range of K values
- Plot the Elbow Curve
- Select and justify the optimal number of clusters based on the plot

b) Silhouette Coefficient
- Calculate silhouette scores across the same range of K values
- Plot the silhouette scores
- Select and justify the optimal K based on the highest silhouette score

2.3 Train and Predict Clusters (25 points):

- Train the K-means model using the selected optimal K
- Predict the cluster label for each customer
- Append the predicted cluster labels as a new column to the dataset

2.4 Visualization of Results (10 points): Create visualizations to display clustering results using the following feature combinations:

- Annual Income (x-axis) vs. Age (y-axis)
- Annual Income (x-axis) vs. Education Level (y-axis) (if available in dataset)

2.5 Interpretation (10 points):

- Provide a detailed description of each identified customer cluster (e.g., high-income, young, educated)
- Propose specific marketing strategies tailored to each customer segment

3. Report and Submission

- Deadline: 11:59 PM, May 10, 2025
- Submission Platform: **Blackboard**
- Submission Format: Submit a single ZIP file named with your student ID, containing the following components:
  - ✓ Source Code (5 points): Clean, well-documented, and executable code
  - ✓ Report (10 points) (PDF format, 2–3 pages):
    - Experimental Setup (5 points): Describe the dataset, preprocessing steps, and clustering methodology
    - Visualization and Interpretation (5 points): Include visualizations from Section 2.4 and insights from Section 2.5
  - ✓ Timeliness (5 points): Late submissions may incur penalties

Ensure that your report is well-structured, concise, and professionally formatted

Good luck!