CA06: Customer Segmentation using K-Means Clustering

Objective: The aim of this assignment is to perform customer segmentation using the K-Means clustering algorithm in order to better understand the different types of customers in a given dataset.

Dataset: You will be working with the 'Mall_Customers.csv' dataset, which can be found at the following link:

https://github.com/ArinB/MSBA-CA-Data/raw/main/CA06/Mall_Customers.csv

The dataset contains the following attributes:

1. CustomerID: Unique ID for each customer
2. Gender: Male or Female
3. Age: Age of the customer
4. Annual Income (k$): Annual income of the customer in thousands of dollars
5. Spending Score (1-100): A score assigned by the mall based on customer behavior and spending nature (higher scores indicate higher spending)

Tasks:

1. Load the dataset and perform exploratory data analysis (EDA): a. Import the necessary libraries (pandas, numpy, matplotlib, seaborn) b. Load the dataset using pandas and display the first few rows c. Check for missing values and handle them appropriately d. Visualize the distribution of features using histograms or boxplots
2. Prepare the data for clustering: a. Perform any necessary feature scaling (StandardScaler or MinMaxScaler) b. Choose the appropriate features for clustering (you may start with 'Annual Income' and 'Spending Score') c. Create a new DataFrame with only the selected features
3. Implement k-means clustering: a. Import the KMeans class from the sklearn.cluster module b. Use the Silhouette Method to determine the optimal number of clusters c. Train the KMeans model with the optimal number of clusters d. Obtain the cluster assignments for each data point
4. Visualize and analyze the clusters: a. Create a scatter plot of the selected features, colored by cluster assignment b. Interpret the clusters and provide a brief

description of each cluster c. (Optional) Perform the same analysis with different sets of features and compare the results

5. Write a report summarizing your findings: a. Describe the dataset and its attributes b. Detail the steps taken for data preprocessing, feature selection, and scaling c. Explain the process of determining the optimal number of clusters d. Describe the clusters and their characteristics e. Discuss any insights or recommendations based on your analysis

Submission: Submit your Jupyter Notebook (or equivalent) containing the code, visualizations, and written analysis of your findings.

Grading Criteria:

- Code quality and correctness (30%)
- Quality of visualizations (20%)
- Clustering analysis and interpretation (30%)
- Report clarity and organization (20%)