Machine Learning - 2nd Assignment

Task 1: Customer Segmentation using K-Means

Name: [Your Name] **Sub Code:** BCS602 **Year:** 2024-2025 **Semester:** VI

---

1. Detailed Explanation of Dataset Used

a. Dataset Name and Source

The dataset used for this task is the **Mall Customer Segmentation Data**. * **Source:** This dataset is a common benchmark dataset, often found on platforms like Kaggle. For this assignment, the `Mall_Customers.csv` file was used. * *(Optional: If you want to cite the specific Kaggle link you might have looked at: e.g., "Originally sourced from Kaggle: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python")*

b. Size and Structure

- **Size:** The dataset contains **200 records (customers)** and **5 attributes**.
- **Attributes:**
  - **CustomerID:** Unique identifier for each customer (Integer). *This column was not used for clustering.*
  - **Gender:** Gender of the customer (Categorical: Male/Female). *This column was not directly used in the primary K-Means clustering for this iteration but could be used for further analysis.*
  - **Age:** Age of the customer (Integer). *While not used in the primary clustering, it was observed in cluster characteristic analysis.*
  - **Annual Income (k$):** Annual income of the customer in thousands of dollars (Integer). **This was a key feature used for clustering.**
  - **Spending Score (1-100):** A score assigned by the mall based on customer behavior and spending nature (Integer, range 1-100). **This was a key feature used for clustering.**

c. Preprocessing of Dataset

The following preprocessing steps were performed: 1. **Feature Selection:** The 'Annual Income (k)'$and'SpendingScore(1-100)'columnswereselectedastheprimaryfeaturesforK-Meansclustering.Thesefeaturesarequantitativeanddirectlyrelevanttocustomersegmentationbasedonpurcha$ $*RenamingColumns : **Foreaseofuseinthecode,'AnnualIncome(k)$' was renamed to `Annual_Income` and 'Spending Score (1-100)' was renamed to `Spending_Score`. 3. **Data Scaling:** The selected features (`Annual_Income` and `Spending_Score`) were scaled using `StandardScaler` from `scikit-learn`. This standardizes features by removing the mean and scaling to unit variance. Scaling is crucial for K-Means as it is a distance-based algorithm and ensures that features with larger magnitudes do not dominate the clustering process.

2. Explain the Working of Algorithm on Selected Dataset with Necessary Figures

a. K-Means Clustering Algorithm

K-Means is an unsupervised machine learning algorithm used for partitioning a dataset into 'k' distinct, non-overlapping clusters. The algorithm works iteratively: 1. **Initialization:** 'k' initial centroids are chosen (e.g., randomly or using 'k-means++'). 2. **Assignment Step:** Each data point is assigned to the nearest centroid, forming 'k' clusters. 3. **Update Step:** The centroid of each cluster is recalculated as the mean of all data points assigned to that cluster. Steps 2 and 3 are repeated until the centroids no longer change significantly or a maximum number of iterations is reached. The objective is to minimize the Within-Cluster Sum of Squares (WCSS), also known as inertia.

b. Determining the Optimal Number of Clusters (k) - Elbow Method

To determine the optimal number of clusters (k) for the K-Means algorithm, the **Elbow Method** was employed. This involves running K-Means for a range of k values (e.g., 1 to 10) and calculating the WCSS (inertia) for each k. A plot of WCSS against k is then generated.

Figure 1: Elbow Method for Optimal k *(Here, you will need to insert the image of your Elbow Method plot from your notebook. You can take a screenshot of the plot from your .ipynb file, save it as an image (e.g.,* `elbow_plot.png`*) in your* `Task1_CustomerSegmentation/report/` *folder, and then reference it in Markdown like this:)* `![Elbow Method Plot](./elbow_plot.png)`

Observation: As seen in Figure 1, the WCSS decreases as k increases. The "elbow" point on the graph, where the rate of decrease sharply changes, indicates a good trade-off between minimizing WCSS and avoiding an excessive number of clusters. For this dataset, the elbow was observed at **k=5**. Therefore, 5 was chosen as the optimal number of clusters.

c. Application of K-Means and PCA for Visualization

Once the optimal k=5 was determined: 1. The K-Means algorithm was applied to the scaled 'Annual_Income' and 'Spending_Score' features with `n_clusters=5`. 2. Each customer was assigned to one of the 5 clusters. 3. **Principal Component Analysis (PCA)** was then used to visualize these clusters. Although the clustering was performed on 2 features, applying PCA (reducing 2 dimensions to 2 principal components) demonstrates the technique as required by the assignment. PCA helps in visualizing high-dimensional data by projecting it onto a lower-dimensional space while retaining the maximum possible variance.

Figure 2: Customer Segments using K-Means (k=5) - Original

Features *(Screenshot of your first scatter plot - Annual Income vs Spending Score, colored by cluster. Save as* `kmeans_original_features_plot.png` *in the* `report` *folder.)* `![K-Means Original Features Plot](./kmeans_original_features_plot.png)`

Figure 3: Customer Segments (PCA-reduced) using K-Means (k=5) *(Screenshot of your PCA scatter plot. Save as* `kmeans_pca_plot.png` *in the* `report` *folder.)* `![K-Means PCA Plot](./kmeans_pca_plot.png)`

The PCA plot (Figure 3) shows the clusters projected onto the first two principal components. The explained variance ratio by PC1 and PC2 was [Insert explained variance ratio from your notebook output, e.g., PC1: 0.55, PC2: 0.45, Total: 1.00, or whatever your notebook shows].

---

3. Interpretation of Tasks and Results

The K-Means clustering algorithm successfully segmented the mall customers into 5 distinct groups based on their annual income and spending scores. The characteristics of these clusters, derived from analyzing their centroids and the distribution of customers (referencing Figure 2 primarily, and supported by PCA visualization in Figure 3), are as follows:

(This is where you expand on the interpretation from your notebook. Be descriptive. Use the cluster numbers as they appear in your plots (e.g., Cluster 0, Cluster 1, ...). You can also refer to the `df.groupby('Cluster').mean()` *output from your notebook.)*

- **Cluster [Number, e.g., 0] - Label: (e.g., "Standard Customers" or "Careful Spenders")**
  - **Characteristics:** These customers generally have [low/moderate/high] annual income and [low/moderate/high] spending scores. (Be specific, e.g., "average annual income around $XXk and spending score around YY").
  - **Potential Business Implication:** [e.g., General marketing, loyalty programs to encourage more spending.]
- **Cluster [Number, e.g., 1] - Label: (e.g., "Target Customers" or "High Value")**
  - **Characteristics:** These customers exhibit [low/moderate/high] annual income and [low/moderate/high] spending scores. (e.g., "high annual income (average $XXk) and high spending scores (average YY)").
  - **Potential Business Implication:** [e.g., Prime targets for premium products, exclusive offers, and personalized marketing.]
- **Cluster [Number, e.g., 2] - Label: (e.g., "Careful Rich" or "High Income, Low Spenders")**
  - **Characteristics:** ...
  - **Potential Business Implication:** ...

- **Cluster [Number, e.g., 3] - Label: (e.g., "Savers" or "Low Income, Low Spenders")**
  - **Characteristics:** ...
  - **Potential Business Implication:** ...
- **Cluster [Number, e.g., 4] - Label: (e.g., "Risk Takers" or "Low Income, High Spenders")**
  - **Characteristics:** ...
  - **Potential Business Implication:** ...

(Make sure the labels and descriptions match YOUR cluster outputs. The order of clusters might be different each time K-Means runs if `random_state` *isn't fixed, or if the K-Means++ initialization leads to a different starting point, though* `random_state=42` *should make it consistent.)*

Conclusion for Task 1: Customer segmentation using K-Means has provided valuable insights into different customer groups. These segments can help the mall management in devising targeted marketing strategies, optimizing store layouts, and improving overall customer experience, ultimately leading to increased sales and customer satisfaction. The use of the Elbow Method helped in scientifically determining the optimal number of segments, and PCA facilitated their visualization.
