Name:-Parth Ulhas Katkar

Roll NO:-282041

PRN:- 22311293

Batch:-B2

**Assignment5:**

**Clustering Customer Data to Identify Profitable Groups in a Mall using Machine Learning**

**1. Problem Statement:**

As a mall owner, it is important to understand customer behavior to improve business strategies and offer personalized services. The dataset includes customer information such as age, gender, annual income, and spending score. Using clustering algorithms, we aim to segment customers into distinct groups to identify the most profitable ones based on their spending behavior.

**2. Dataset Description:**

**Source:** [Kaggle - Mall Customer Segmentation](https://www.kaggle.com/shwetabh123/mall-customers)

**Attributes:**

* CustomerID
* Gender
* Age
* Annual Income (k$)
* Spending Score (1-100)

**3. Methodology:**

**a) Data Pre-processing:**

* Load the dataset using Pandas
* Check for null/missing values
* Convert categorical features (like Gender) to numerical format

**b) Data Preparation:**

* Select relevant features: ['Annual Income (k$)', 'Spending Score (1-100)']
* No need for train-test split, as clustering is unsupervised

**c) Apply Machine Learning Algorithms:**

* **K-Means Clustering**
  + Use Elbow Method to find optimal number of clusters
* **Hierarchical Clustering**
  + Create dendrogram to decide number of clusters
  + Apply Agglomerative Clustering

**4. Python Implementation:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import LabelEncoder

from scipy.cluster.hierarchy import dendrogram, linkage

from sklearn.cluster import AgglomerativeClustering

# Load dataset

df = pd.read\_csv("Mall\_Customers.csv")

# Data Preprocessing

df.isnull().sum() # Check for missing values

# Encode Gender

le = LabelEncoder()

df['Gender'] = le.fit\_transform(df['Gender'])

# Selecting relevant features

X = df[['Annual Income (k$)', 'Spending Score (1-100)']]

# 1. K-Means Clustering

wcss = []

for i in range(1, 11):

km = KMeans(n\_clusters=i, random\_state=42)

km.fit(X)

wcss.append(km.inertia\_)

# Elbow Method Plot

plt.plot(range(1, 11), wcss, marker='o')

plt.title('Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel('WCSS')

plt.show()

# From the Elbow method, assume optimal k = 5

kmeans = KMeans(n\_clusters=5, random\_state=42)

df['KMeans\_Cluster'] = kmeans.fit\_predict(X)

# 2. Hierarchical Clustering

linked = linkage(X, method='ward')

plt.figure(figsize=(10, 7))

dendrogram(linked)

plt.title('Dendrogram')

plt.show()

# Apply Agglomerative Clustering

hc = AgglomerativeClustering(n\_clusters=5)

df['Hierarchical\_Cluster'] = hc.fit\_predict(X)

# Visualize Clusters

plt.figure(figsize=(10, 5))

sns.scatterplot(x='Annual Income (k$)', y='Spending Score (1-100)',

hue=df['KMeans\_Cluster'], palette='Set1')

plt.title('Customer Segmentation (K-Means)')

plt.show()

**5. Results and Observations:**

* From **K-Means** and **Hierarchical Clustering**, the data was best divided into **5 clusters**.
* One cluster contains customers with **high income and high spending score** — the most profitable group.
* Another cluster includes **low income and low spenders** — not profitable.

**6. Conclusion:**

By applying **K-Means** and **Hierarchical Clustering**, we effectively segmented mall customers into 5 distinct groups. This segmentation helps in:

* Targeting **high-value customers** with special offers
* Improving **marketing strategies**
* Enhancing **customer satisfaction** by personalizing services