**Name: Aditee Mohanty**

**Class: CSE(AI) B div**

**Roll no.: 282023**

**PRN No.: 22310631**

**Subject: Machine Learning**

**Assignment 5 – Customer Segmentation Using Clustering**

**Objective**

The goal of this assignment is to identify profitable customer groups in a mall based on their **Annual Income** and **Spending Score** using **Clustering algorithms**. This helps the mall owner make informed marketing decisions and offer targeted promotions.

**Dataset Used**

**Mall Customers Dataset** from Kaggle:  
<https://www.kaggle.com/shwetabh123/mall-customers>

**Attributes in the dataset:**

* **CustomerID**: Unique ID for each customer
* **Gender**: Male/Female
* **Age**
* **Annual Income (k$)**
* **Spending Score (1–100)**

**a) Data Preprocessing**

* The dataset was checked for **null values** and **data types** using .info() and .isnull().sum().
* The **‘Gender’** column was converted to numeric using **Label Encoding**.
* Only the relevant features – **Annual Income** and **Spending Score** – were selected for clustering.
* Features were **standardized** using StandardScaler to bring them to a similar scale for better clustering performance.

**b) Data Preparation**

* The dataset was split into **training and testing sets** using an 80-20 ratio with train\_test\_split.

**c) Machine Learning Algorithms Applied**

**1. K-Means Clustering**

* The optimal number of clusters (**K**) was identified using the **Elbow Method**, with WCSS plotted for K=2 to K=10.
* Based on the Elbow curve, **K=5** was chosen and applied.
* Cluster labels were stored in the dataset for visualization.

**2. Agglomerative Clustering**

* Another unsupervised algorithm, **Agglomerative Hierarchical Clustering**, was also applied with **5 clusters**.
* Cluster labels from Agglomerative Clustering were also stored for comparison.

**d) Model Evaluation**

The **Silhouette Score** was used to evaluate the quality of clustering for both algorithms:

* **K-Means Silhouette Score**: (Printed in output)
* **Agglomerative Clustering Silhouette Score**: (Printed in output)

Higher silhouette scores indicate better-defined clusters.

**e) Cross-Validation**

Although clustering is unsupervised, cross-validation techniques were demonstrated to show model reliability using dummy data:

**Cross-Validation Techniques Explored:**

* **K-Fold**
* **Leave-One-Out**
* **Leave-P-Out**
* **Stratified K-Fold**
* **Repeated K-Fold**

These techniques were shown with small example datasets to illustrate how training and test splits are made during validation processes.

**Cluster Visualization**

The K-Means clusters were visualized using a **scatter plot** with:

* X-axis: **Annual Income**
* Y-axis: **Spending Score**
* Different colors representing different customer segments

This helped in identifying profitable customer groups (e.g., high income, high spenders).

**Conclusion**

Clustering allowed us to segment mall customers based on their spending behavior and income. The two algorithms, **K-Means** and **Agglomerative Clustering**, successfully grouped customers into distinct segments. These insights can help the mall owner in targeting the right audience for special offers and personalized services.