**Project Documentation: Mall Customer Segmentation Using Unsupervised Machine Learning**

1. **Introduction**

* **Objective**: To segment mall customers into distinct groups based on spending behavior, income, and demographics using unsupervised learning techniques.
* **Business Value:** Enables targeted marketing, inventory optimization, and personalized customer experiences.

2. **Dataset Overview**

* **Source**: [Mall Customer Data](https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python)

**Features**

- CustomerID (Unique identifier)

- Gender (Male/Female)

- Age

- Annual Income (k$)

- Spending Score (1-100)

* **Size**: 200 customers, 5 features

3. **Methodology**

A. **Data Preprocessing**

1. Handled Missing Values: No missing data found.

2. Feature Engineering:

- Dropped CustomerID (irrelevant for clustering).

- Encoded Gender (Male: 0, Female: 1).

3. Scaling: Standardized features using StandardScaler.

B. **Clustering Techniques**

1. K-Means Clustering:

- Optimal clusters (k=5) determined via the Elbow Method.

- Silhouette Score: 0.45 (moderate separation).

2. Alternative Algorithms (Optional):

- DBSCAN (for density-based clustering).

- Hierarchical Clustering (for dendrogram visualization).

C. **Evaluation Metrics**

- Within-Cluster Sum of Squares (WCSS): Used for Elbow Method.

- Silhouette Analysis: Validated cluster cohesion/separation.

4. **Results & Insights**

**Identified Clusters**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Profile** | **Key Characteristics** | **Business Strategy** |
| 0 | High Income, Low Spending | Older, cautious spenders | Luxury items with value messaging |
| 1 | Moderate Income & Spending | Middle-aged, balanced | Loyalty programs |
| 2 | High Income, High Spending | Young affluent shoppers | Premium products/VIP services |
| 3 | Low Income, High Spending | Young, spendthrift | Trendy items with payment plans |
| 4 | Low Income, Low Spending | Budget-conscious | Discounts and promotions |

**Visualizations**

1. 2D Scatter Plot: Income vs. Spending Score with clusters.

2. Box Plots: Age distribution per cluster.

3. Cluster Centroids: Mean values of each feature per group.

5**. Technical Implementation**

**Code Snippets**

A. **Data Preprocessing**

##Standardize features

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(df[['Age', 'Annual Income', 'Spending Score']])

B. **K-Means Clustering**

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

clusters = kmeans.fit\_predict(X\_scaled)

C. **Visualization**

plt.scatter(X\_scaled[:,1], X\_scaled[:,2], c=clusters, cmap='viridis')

plt.xlabel('Annual Income')

plt.ylabel('Spending Score')

6. **Discussion**

**Strengths**

- Automated customer segmentation without predefined labels.

- Actionable insights for marketing teams.

**Limitations**

- Assumes spherical clusters (K-Means limitation).

- Requires manual interpretation of clusters.

**Future Work**

1. 3D Clustering: Add Age as a third dimension.

2. Real-Time Segmentation: Deploy as a web app using Streamlit.

3. Algorithm Comparison: Test DBSCAN/Hierarchical clustering.

7. **Conclusion**  :This project successfully demonstrated how unsupervised learning can extract meaningful customer segments from raw transactional data, providing retailers with a competitive edge through data-driven decision-making.