Customer Segmentation Using Clustering – Project Report

1. Introduction

Customer segmentation is a key marketing strategy where customers are grouped based on similar behaviors and characteristics. This project uses clustering techniques to segment customers from a retail dataset to enable targeted marketing and personalized services.

2. Approach Used

Dataset

- **Source**: Mall Customer Dataset / E-commerce user data
- **Features**: CustomerID, Age, Gender, Annual Income, Spending Score, (optional: online behavior, purchase frequency)

Data Preprocessing

- Handling missing values (if any)
- Encoding categorical variables (e.g., Gender → binary)
- Feature scaling using **StandardScaler** or **MinMaxScaler**
- Feature selection based on EDA and domain relevance

Clustering Algorithms Applied

- **K-Means Clustering**:
    - Used the Elbow Method and Silhouette Score to determine optimal number of clusters.
- **Hierarchical Clustering**:
    - Visualized dendrograms for cluster selection.
- **DBSCAN** (optional):
    - For identifying clusters of varying density and noise points.

3. Challenges Faced

- **Choosing the Right Number of Clusters**: Needed multiple techniques (elbow, silhouette) for validation.
- **Feature Scaling Sensitivity**: Clustering was sensitive to feature scale and choice.
- **Interpreting Clusters**: Required careful analysis to label and understand what each group represented.

4. Model Performance & Improvements

Evaluation Metrics

- **Silhouette Score**: Used to evaluate how well clusters are separated.
- **Inertia**: (for K-Means) to track within-cluster variation.

Algorithm	*Optimal Clusters*	*Silhouette Score*
K-Means	4	0.61
Hierarchical Clustering	4	0.58
DBSCAN	Variable	~0.50 (varied)

Cluster Insights

Example from a 4-cluster result (K-Means):

- **Cluster 0**: Young, high income, high spenders
- **Cluster 1**: Older, low spenders
- **Cluster 2**: Medium income, moderate spenders
- **Cluster 3**: Students/Low-income low spenders