REPORT
Task 02: Customer Segmentation using K-Means Clustering

1. Introduction

Customer segmentation is one of the most important applications of Machine Learning in the retail industry. It helps businesses identify different groups of customers based on their purchasing behavior and spending patterns. By understanding customer groups, companies can design targeted marketing strategies, improve customer retention, and increase revenue.

In this project, **K-Means Clustering**, an unsupervised learning algorithm, is used to group customers of a retail store based on purchase behavior-related features. The final output is a set of customer clusters representing different types of customers.

2. Objective

The objective of this task is to create a **K-Means clustering model** to group customers into different segments based on their purchase behavior.

The model aims to:
• Identify customer groups based on spending patterns
• Help businesses understand customer categories
• Support targeted promotions and marketing strategies

3. Dataset Description

The dataset used for this task is the **Mall Customers Dataset** (commonly used for customer segmentation tasks). It contains customer information such as Customer ID, Gender, Age, Annual Income, and Spending Score.

For clustering, the following features were selected:
• **Annual Income (k$)**
• **Spending Score (1–100)**

These features are suitable for grouping customers into meaningful categories.

4. Methodology

This project follows the standard Machine Learning pipeline for clustering:

4.1 Data Loading: Loaded the dataset using Pandas and checked shape, data types, and missing values.
4.2 Feature Selection: Selected Annual Income and Spending Score for clustering.
4.3 Feature Scaling: Used StandardScaler to normalize features for distance-based clustering.
4.4 Finding Optimal K: Used Elbow Method and Silhouette Score to determine a suitable number of clusters.
4.5 Model Training: Trained K-Means and assigned cluster labels to customers.
4.6 Visualization: Plotted clusters to visually analyze segmentation.

5. Results & Interpretation

After applying clustering, customers were grouped into multiple segments. Typical groups include:
• High Income + High Spending → Premium customers
• High Income + Low Spending → Potential customers
• Low Income + High Spending → Deal seekers
• Low Income + Low Spending → Budget customers
• Average Income + Average Spending → Regular customers

A cluster summary table was generated by calculating mean values for each cluster to support interpretation.

6. Output

The notebook generates:
• Elbow Curve (WCSS vs K)
• Silhouette Score Plot
• Scatter plot showing customer clusters
• Cluster summary table
• Exported segmented dataset as **clustered_customers.csv**

7. Conclusion

This project successfully implemented **K-Means clustering** to group retail customers into segments based on income and spending behavior. The results provide insights that can help retailers with marketing, customer targeting, and business growth strategies.

8. Future Improvements

Improvements can include:
• Including more features such as Age and Gender
• Using real purchase history data (frequency, total spend)
• Trying other clustering algorithms like DBSCAN or Hierarchical Clustering
• Implementing advanced RFM segmentation
• Deploying as an interactive dashboard using Streamlit

9. Tools and Libraries Used

Python, Pandas, NumPy, Matplotlib, Seaborn, scikit-learn (KMeans, StandardScaler, silhouette_score), Jupyter Notebook

Author

Pratham Agarwal
Machine Learning Intern – Prodigy InfoTech
GitHub: https://github.com/1234pratham2k6k1234-glitch