Customer Segmentation using Unsupervised Machine Learning

Project Overview
This project applies unsupervised machine learning techniques to segment customers based on behavioral and demographic attributes. The objective is to discover hidden patterns in unlabeled data using K-Means and DBSCAN clustering algorithms and compare their performance.

Objectives
• Perform data preprocessing on unlabeled data
• Apply K-Means and DBSCAN clustering algorithms
• Determine optimal clusters using the Elbow Method
• Evaluate clustering quality using Silhouette Score
• Visualize clusters using PCA

Dataset
Source: Kaggle – Mall Customer Segmentation Dataset
Features: Gender, Age, Annual Income, Spending Score

Technologies Used
Python, NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn

Machine Learning Pipeline
1. Data Preprocessing – Cleaning, handling missing values
2. Encoding – Label Encoding for categorical data
3. Feature Scaling – StandardScaler normalization
4. Model Training – K-Means and DBSCAN
5. Dimensionality Reduction – PCA for visualization

Algorithms
K-Means: Centroid-based clustering requiring predefined cluster count.
DBSCAN: Density-based clustering detecting noise and arbitrary shaped clusters.

Performance Metrics
Silhouette Score used to evaluate clustering quality for both algorithms.

Results & Observations
K-Means produced well-defined clusters with higher silhouette score. DBSCAN effectively detected noise but formed fewer clusters.

Conclusion
K-Means was more suitable for this dataset due to structured cluster distribution. DBSCAN provided insights into noise and density behavior.

Future Enhancements
• Hierarchical Clustering
• DBSCAN parameter tuning
• Larger datasets
• Visualization dashboards

Author
Kavindu Hirushan
Electrical & Electronics Engineering Undergraduate