Green University of Bangladesh
Department of Computer Science and Engineering (CSE)

Faculty of Sciences and Engineering

Semester: (Spring, Year: 2025), B.Sc. in CSE (Day)

LAB REPORT NO 08

Course Title: Data Mining Lab

Course Code: CSE 436 Section: 213 D4

Lab Experiment Name:

"Clustering Algorithms: Hierarchical and Density-Based Approaches on Real-Life Customer Data"

Student Details

	Name	ID
1.	Md. Rabby Khan	213902037

Lab Date : 18/04/2025

Submission Date : 25/04/2025

Course Teacher's Name : Md. Jahid Tanvir

[For Teachers use only: Don't Write Anything inside this box]

Lab Report Status

Marks: …… Signature: .....

Comments: ..... Date: .....

1. TITLE OF THE LAB EXPERIMENT

Implementation of Hierarchical and Density-Based Clustering Algorithms on Real-Life Customer Data

2. OBJECTIVES/AIM

- To apply clustering techniques (Hierarchical Clustering and DBSCAN) to group customers based on their data.
- To visualize customer distribution using scatter plots of income and spending scores.
- To create a dendrogram with Hierarchical Clustering to find the best number of clusters.
- To use DBSCAN to identify clusters and detect outliers.
- To analyze the clustering results to understand how customers are grouped based on income and spending behavior.

3. PROCEDURE / DESIGN

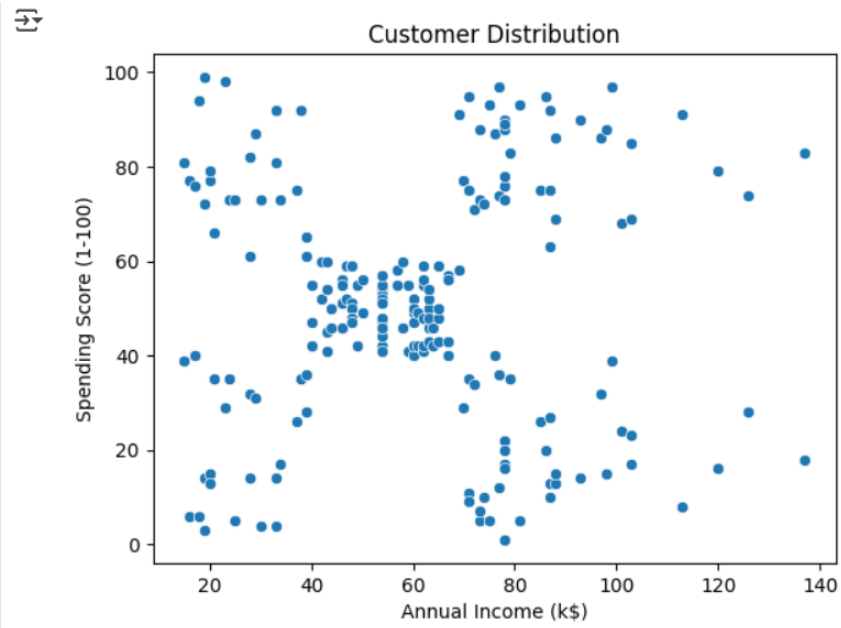
- **Load and Preprocess Data**: Import the dataset and select features Annual Income and Spending Score. Standardize the data.
- **Visualize Data:** Create a scatter plot to display customer distribution.
- **Hierarchical Clustering**: Apply Agglomerative Clustering and visualize the clusters.
- **Generate Dendrogram:** Plot a dendrogram to determine the optimal number of clusters.
- **DBSCAN Clustering:** Apply DBSCAN and visualize the clustering results.
- **Compare Results:** Compare the outcomes of Agglomerative-Clustering and DBSCAN.
- **Interpret and Derive Insights:** Analyze the clusters for customer segmentation and marketing insights.

4. IMPLEMENTATION/ CONFIGURATION

Step 1: Load and Preprocess Dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("/content/sample_data/Mall_Customers.csv")

X = df[['Annual Income (k$)', 'Spending Score (1-100)']].values

sns.scatterplot(x=X[:, 0], y=X[:, 1])
plt.xlabel("Annual Income (k$)")
plt.ylabel("Spending Score (1-100)")
plt.title("Customer Distribution")
plt.show()
```

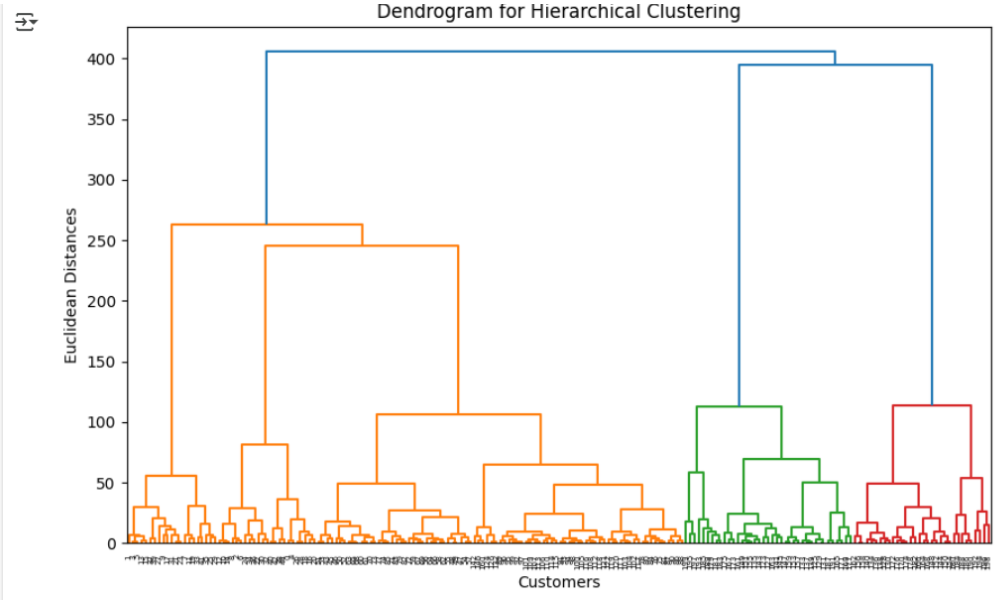
5. TEST RESULT / OUTPUT 1

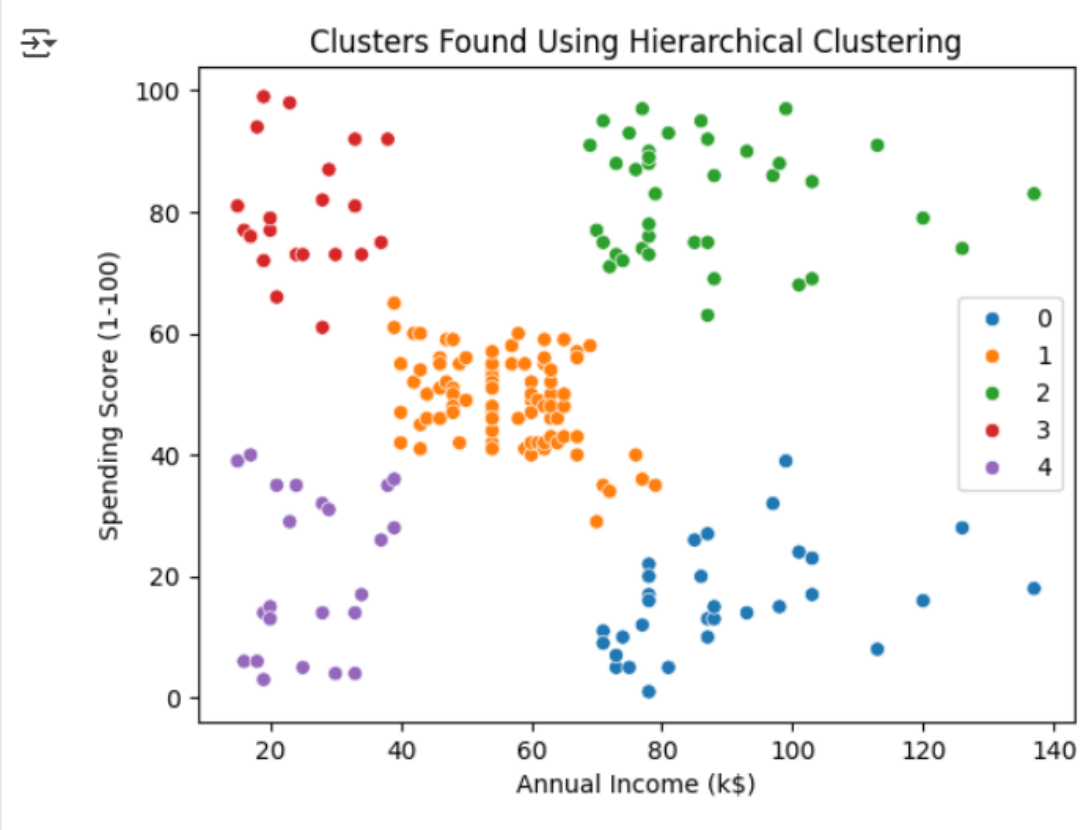


Step 2: Hierarchical Clustering

```
import scipy.cluster.hierarchy as shc
from sklearn.cluster import AgglomerativeClustering

plt.figure(figsize=(10, 6))
plt.title("Dendrogram for Hierarchical Clustering")
dendrogram = shc.dendrogram(shc.linkage(X, method='ward'))
plt.xlabel("Customers")
plt.ylabel("Euclidean Distances")
plt.show()

hc = AgglomerativeClustering(n_clusters=5)
y_hc = hc.fit_predict(X)

sns.scatterplot(x=X[:, 0], y=X[:, 1], hue=y_hc, palette="tab10")
plt.title("Clusters Found Using Hierarchical Clustering")
plt.xlabel("Annual Income (k$)")
plt.ylabel("Spending Score (1-100)")
plt.show()
```

5. TEST RESULT / OUTPUT 2

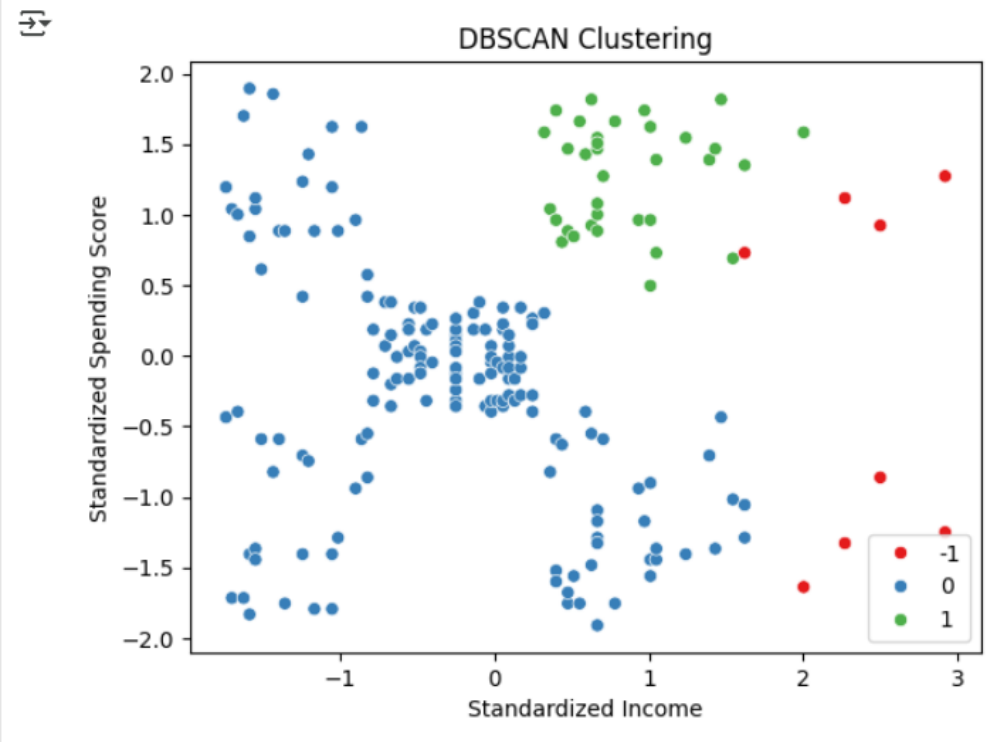




Step 3: Density-Based Clustering (DBSCAN)

```
from sklearn.cluster import DBSCAN
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

db = DBSCAN(eps=0.5, min_samples=5)
y_db = db.fit_predict(X_scaled)

sns.scatterplot(x=X_scaled[:, 0], y=X_scaled[:, 1], hue=y_db, palette="Set1")
plt.title("DBSCAN Clustering")
plt.xlabel("Standardized Income")
plt.ylabel("Standardized Spending Score")
plt.show()
```

5. TEST RESULT / OUTPUT 3



EXPLANATION: The scatter plot shows the relationship between annual income and spending scores, helping identify customer distribution. The dendrogram from hierarchical clustering identifies 5 distinct customer groups based on income and spending patterns. DBSCAN clustering detects dense clusters and outliers, highlighting unusual customer behaviors.

6. ANALYSIS AND DISCUSSION

In this experiment, I applied two clustering techniques Hierarchical Clustering and DBSCAN to segment customers based on their income and spending scores. Hierarchical Clustering allowed us to visualize customer groups through a dendrogram, from which we identified five distinct clusters. These clusters represented different customer profiles, such as high-income, high-spending individuals and low-income, low-spending individuals.

On the other hand, DBSCAN, a density-based clustering method, grouped customers by identifying dense regions and marking sparse points as outliers. This approach revealed irregular customer patterns and detected outliers that were not captured by Hierarchical Clustering.

When comparing the two methods, Hierarchical Clustering provided a clear division of customers into well-defined groups, making it suitable for applications that require explicit segmentation. However, DBSCAN was more flexible, capable of detecting non-spherical clusters and identifying outliers. Its ability to recognize noise and unusual patterns in the data highlights DBSCAN's effectiveness in detecting customer behaviors that deviate from the norm.

In conclusion, both clustering methods were effective in segmenting the data. Hierarchical Clustering offered clearer groupings, while DBSCAN uncovered hidden patterns and outliers. These insights can support businesses in customer profiling, targeted marketing, and anomaly detection.

7. SUMMARY

In this lab experiment, I applied Hierarchical Clustering and DBSCAN to segment customers based on their annual income and spending scores. Hierarchical Clustering provided clear, well-defined customer groups, with 5 clusters identified through the dendrogram. In contrast, DBSCAN identified clusters based on density and flagged outliers, offering more flexibility and revealing hidden patterns in the data. Both methods provided valuable insights, with Hierarchical Clustering offering explicit segmentation and DBSCAN detecting anomalies. The results can be used for customer profiling, targeted marketing, and identifying irregular behaviors. This experiment demonstrated the effectiveness of both clustering techniques in real-life data analysis.