Green University Of Bangladesh

Department Of Computer Science and Engineering (CSE)

Faculty of Sciences and Engineering

Semester: (Fall, Year: 2023), B.Sc. in CSE (DAY)

LAB REPORT NO - 06

Course Title: Data Mining Lab

Course Code: CSE-436 **Section:** D2

Lab Experiment Name: Implement hierarchical and density-based clustering

Student Details

	Name	ID
1	Sk. Nahid	201902073

Lab Date : 02/12/2023

Submission Date : 08/12/2023

Course Teacher Name : **Rezwanul Haque**

Lab Report Status

Mark:..... **Signature:.....**

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

1 INTRODUCTION

Clustering (or cluster analysis) is a technique that allows us to find groups of similar objects that are more related to each other than to objects in other groups.

2 OBJECTIVE

This lab report aims to determine different clustering methods, like hierarchical and density-based clustering.

3 PROCEDURE

The penguins dataset was initially loaded and preprocessed by selecting relevant features—bill length and flipper length—followed by removal of any missing values. Two distinct clustering techniques were then applied. Firstly, hierarchical clustering was performed using the linkage algorithm with the ward method, generating a dendrogram that visually depicted the hierarchical structure of clusters. Simultaneously, density-based clustering was implemented with the DBSCAN algorithm after standardizing the features. This approach, not requiring a predetermined number of clusters, identified clusters based on data density, accommodating irregularly shaped groups and detecting outliers. The combination of hierarchical and density-based clustering provided a holistic exploration of inherent structures in the dataset, offering insights into both hierarchical relationships and density-driven cluster formations.

4 IMPLEMENTATION

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import AgglomerativeClustering
from sklearn.preprocessing import StandardScaler
from scipy.cluster import hierarchy
from sklearn.cluster import DBSCAN
df = pd.read_csv('/kaggle/input/penguins/penguins.csv')
print(df.shape) # (344, 9)
df = df[['bill_length_mm', 'flipper_length_mm']]
df = df.dropna(axis=0)
```

Listing 1: Import Library & Dataset

```
clusters = hierarchy.linkage(df, method="ward")

plt.figure(figsize=(8, 6))
dendrogram = hierarchy.dendrogram(clusters)
```

```
# Plotting a horizontal line based on the first biggest distance between
    clusters
plt.axhline(150, color='red', linestyle='--');
# Plotting a horizontal line based on the second biggest distance between
    clusters
plt.axhline(100, color='crimson');
```

Listing 2: Hierarchical Clustering

```
scaler = StandardScaler()

df_scaled = scaler.fit_transform(df)
dbscan_cluster = DBSCAN(eps=0.5, min_samples=5)
df['DBSCAN_Cluster'] = dbscan_cluster.fit_predict(df_scaled)

plt.scatter(df_scaled[:, 0], df_scaled[:, 1], c=df['DBSCAN_Cluster'], cmap
    ='viridis')
plt.title('DBSCAN Clustering')
plt.xlabel('Standardized Bill Length (mm)')
plt.ylabel('Standardized Flipper Length (mm)')
plt.show()
```

Listing 3: Density Based Clustering

5 OUTPUT

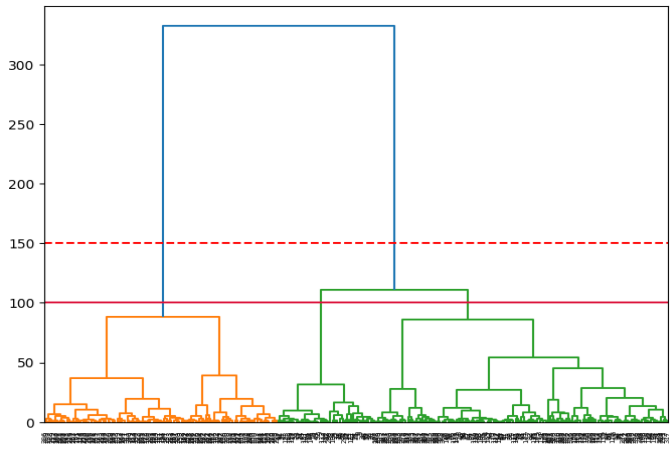

Figure 1: Dataset details

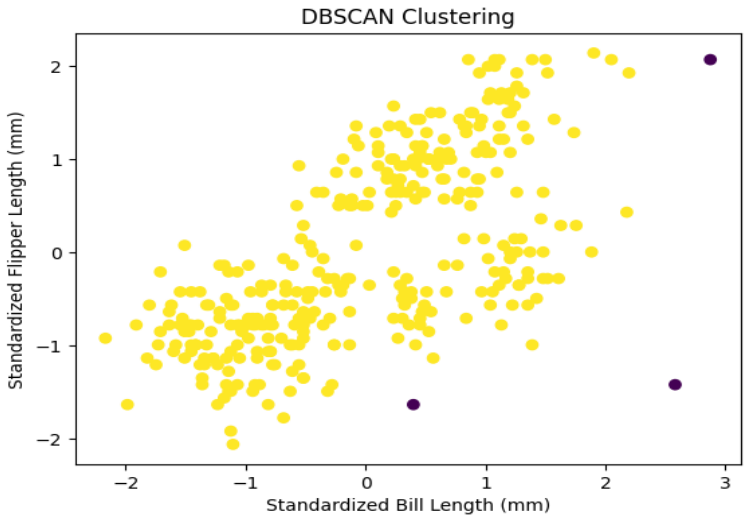

Figure 2: Dataset details

6 DISCUSSION & ANALYSIS

The application of hierarchical and density-based clustering techniques on the penguins dataset revealed hierarchical relationships and density-driven clusters, offering a nuanced perspective on inherent structures in the data. This dual approach enhanced interpretability by capturing both hierarchical patterns and accommodating irregularly shaped clusters without predetermined counts.