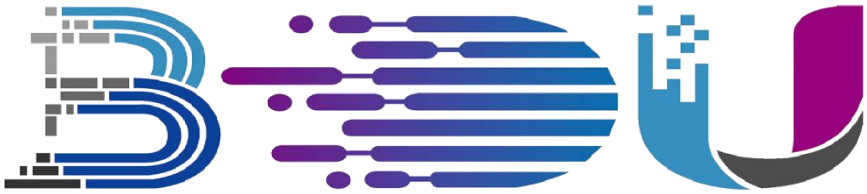

Bangabandhu Sheikh Mujibur Rahman
Digital University

Course Title: Data Science

Course Code: IOT4313

Assignment No: 02 (Clustering)

SUBMITTED To:

Nurjahan Nipa
Lecturer,
Department of IRE, BDU.

Submitted By:

Sieadatun Naher Kanon

ID: 1901018

Session: 2019-20

Date of Submission: 14/10/2023.

Assignment on Clustering

Introduction: Clustering is the task of dividing the unlabeled data or data points into different clusters such that similar data points fall in the same cluster than those which differ from the others.

Here is a detailed explanation of the clustering algorithms required in Parts A, B, and C.

Part A: K-Means Clustering

K-Means clustering is a type of unsupervised machine learning algorithm used to partition a set of data points into distinct groups or clusters based on their similarity.

Approach:

- Here, we used data from the "Mall_Customer" datasets and focused on specific customer attributes: Age, Annual Income, and Spending Score. We wanted to group customers based on similarities.
- We applied the K-Means clustering algorithm, testing different cluster counts from 1 to 15. For each test, we measured how much data points spread within clusters, called Sum of Squared Errors (SSE).

- The goal was to find the ideal number of clusters. We plotted the SSE values against the number of clusters to identify a point where the SSE change slows down, which suggests the right number of clusters

Results:

• According to the Elbow Method, the ideal number of clusters is determined.

• Using K-Means, customers have been categorized into five distinct clusters based on shared characteristics related to age, annual income, and spending score.

• These cluster groupings offer valuable insights that can inform marketing and business strategies.

Part B: Hierarchical Clustering

Hierarchical clustering is a type of unsupervised machine learning algorithm used for grouping similar data points into clusters or hierarchical structures.

Results:

• Hierarchical Clustering generated a dendrogram that visually illustrates the hierarchical connections between clusters.

• By selecting a specific level to cut the dendrogram, we can establish the desired number of clusters.

• This approach is valuable for gaining insights into the hierarchical organization of the data.

Part C: Density-Based Clustering

Density-Based Clustering is a type of unsupervised machine learning algorithm used to identify clusters in data with varying densities. Unlike traditional methods like K-Means, which assume that clusters are roughly spherical and equally sized, density-based clustering algorithms discover clusters of arbitrary shapes and sizes based on the density of data points.

Results:

- DBSCAN grouped data points according to their local density, forming clusters with diverse shapes and sizes.
- Anomalies were designated as noise and marked with the label (-1).
- The level of cluster detail could be regulated by modifying the epsilon (ε) value.
- This method is particularly effective for identifying clusters that do not conform to standard shapes.

https://github.com/1901018/-DS_Assignment_2/upload/main