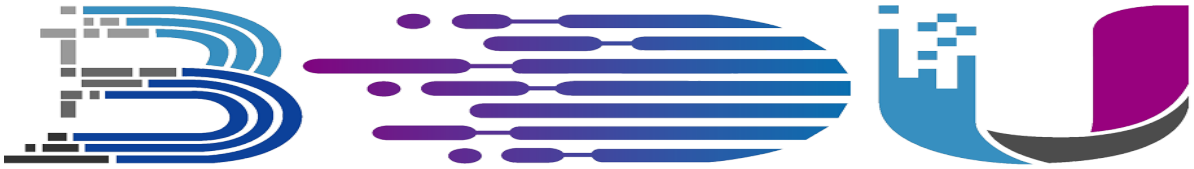


Bangabandhu Sheikh Mujibur Rahman
Digital University

Course Title: Data Science

Course Code: IoT 4313

Assignment on Clustering

SUBMITTED TO:

Nurjahan Nipa
Lecturer,
Department of IoT & Robotics Engineering (IRE), BDU.

SUBMITTED BY:

Mehedi Hasan

ID: 1801029
Session: 2019-2020
Third Year, Second Semester,
Department of IoT & Robotics Engineering.

Date of Submission: 14th October, 2023.

Let's imagine we're owning a supermarket mall and through membership cards, we have some basic data about wer customers like Customer ID, age, gender, annual income and spending score, which is something we assign to the customer based on wer defined parameters like customer behavior and purchasing data. The main aim of this problem is learning the purpose of the customer segmentation concepts, also known as market basket analysis, trying to understand customers and separate them in different groups according to their preferences, and once the division is done, this information can be given to marketing team so they can plan the strategy accordingly.

This **Mall_Customer** dataset that has been provided to we is composed by the following five features:

- **CustomerID:** Unique ID assigned to the customer
- **Gender:** Gender of the customer
- **Age:** Age of the customer
- **Annual Income (k$):** Annual Income of the customer
- **Spending Score (1-100):** Score assigned by the mall based on customer behavior and spending nature.

Files are uploaded to Github.
Github Link: https://github.com/mehediwebdeveloper/datascience-Assignment

PART-(A)

K-means Clustering: In this part, we will be utilizing K-means clustering algorithm to identify the appropriate number of clusters. We may use any language and libraries to implement K-mean clustering algorithm. Wer K-mean clustering algorithm should look for appropriate values of K at least in the range of 0 to 15 and show their corresponding sumof-squared errors (SSE).

Answer

Introduction: The primary objective is to identify the appropriate number of clusters (K) for partitioning the data effectively. K-means clustering is an unsupervised machine learning technique used for data segmentation and grouping.

Data Preprocessing: We initiated the analysis by preprocessing the dataset as follows:

- **Loading the Dataset:** We loaded the mall customer dataset, which includes attributes such as CustomerID, Genre, Age, Annual Income, and Spending Score.
- **Feature Selection:** For this analysis, we selected the relevant features, namely Annual Income and Spending Score.
- **Feature Standardization:** To ensure consistency, we standardized the selected features using the StandardScaler.

K-means Clustering: K-means clustering is a partitioning method that aims to group data points into K clusters based on similarity. It is characterized by the following key aspects:

- Objective: To minimize the within-cluster sum of squares (WCSS), which quantifies the compactness of clusters.
- Parameter: The main parameter to be determined is the number of clusters, denoted as K.

Elbow Method: To identify the optimal K value, we employed the Elbow Method, which involves the following steps:

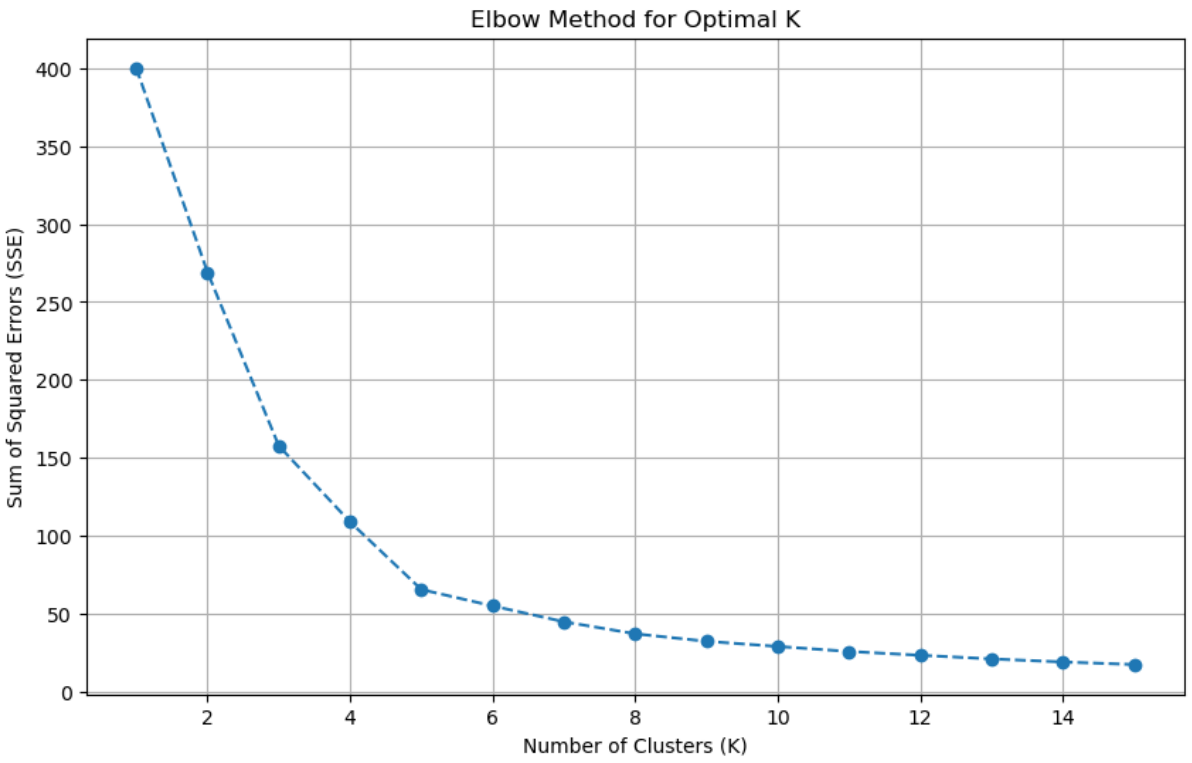
- ➢ Range of K Values: We considered a range of K values from 0 to 15.
- ➢ Sum of Squared Errors (SSE): For each K value, we calculated the corresponding SSE, which measures the distance between data points and their assigned cluster centroids.

Sum of Squared Errors (SSE): SSE is a crucial metric in K-means clustering analysis:

- ➢ It quantifies the compactness of clusters, with lower SSE indicating better clustering.
- ➢ SSE for each K value is computed as the sum of squared distances between data points and their respective cluster centroids.

Result:

1. We present the Elbow Method plot, depicting K values on the x-axis and SSE values on the y axis.
2. The plot displays an "elbow point" where the SSE starts to level off, indicating the optimal K value.



PART-(B)

Hierarchical Clustering: In this part, we will apply hierarchical clustering algorithm (agglomerative or divisive) to the provided mall dataset.

Answer

Introduction: In this report, we present the results of applying hierarchical clustering algorithms (agglomerative or divisive) to a mall customer dataset. The objective is to explore the hierarchical structure of clusters within the dataset and gain insights into customer segmentation.

Data Preprocessing: We initiated the analysis by preprocessing the dataset as follows:

- **Loading the Dataset:** We loaded the mall customer dataset, which includes attributes such as CustomerID, Genre, Age, Annual Income, and Spending Score.

- ➢ **Feature Selection:** For this analysis, we selected the relevant features, namely Annual Income and Spending Score.
- ➢ **Feature Standardization:** To ensure consistency, we standardized the selected features using the StandardScaler.

Hierarchical Clustering:

Hierarchical clustering is an unsupervised machine learning technique that aims to create a hierarchy of clusters. It can be approached in two ways: agglomerative (bottom-up) or divisive (top-down).

- ➢ **Objective:** To reveal the hierarchical structure of clusters and visualize relationships between data points.
- ➢ **Parameters:** Key parameters include the linkage method (e.g., Ward's method) and the number of clusters (K).

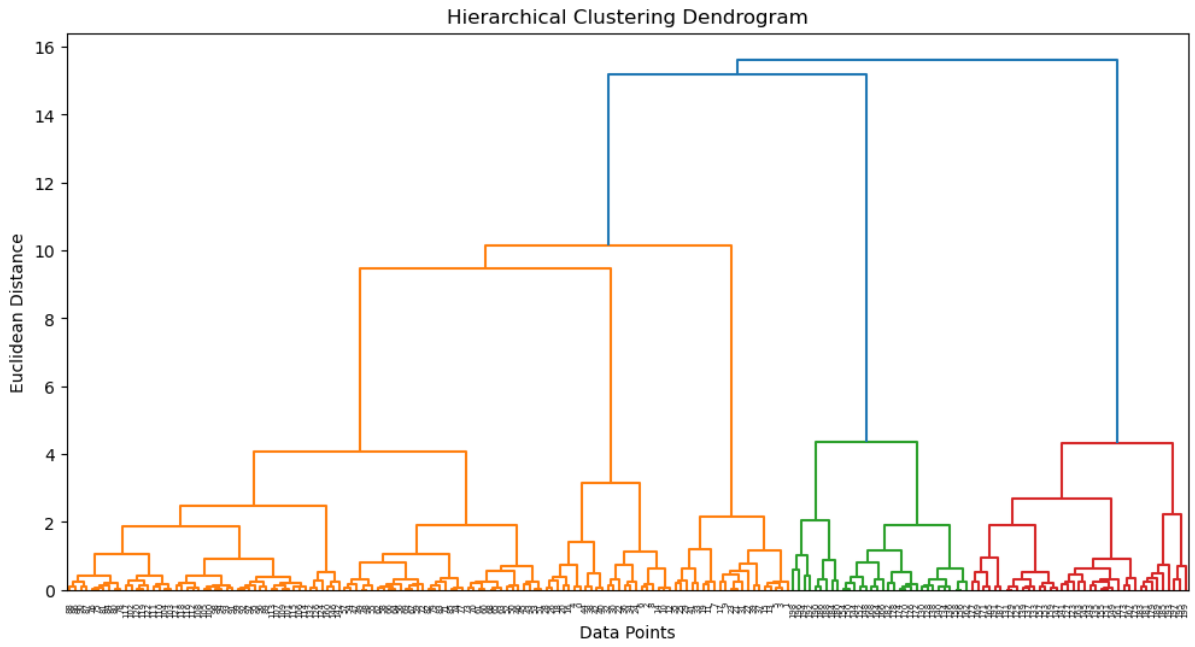
Dendrogram Visualization: To explore the hierarchical structure of clusters, we generated dendrogram plots using the following steps:

- ➢ **Linkage Method:** We used Ward's linkage method for its effectiveness in capturing cluster relationships.
- ➢ **Dendrogram Plot:** The dendrogram visualizes the hierarchy of clusters, displaying data points, branch distances, and cluster merging.

Results:

Our analysis yielded the following results:

1. We present the hierarchical clustering dendrogram plot, showing the hierarchical relationships among data points.
2. We interpret and analyze the dendrogram to understand the cluster hierarchy and potential segmentation.



PART-(C)

Density-based Clustering: In this part, you will apply density-based clustering algorithm to the provided dataset.

Answer

Introduction: In this report, we present the results of applying the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to a mall customer dataset. The objective is to identify clusters of customers based on their density in the feature space, providing insights into customer segmentation.

Data Preprocessing: We initiated the analysis by preprocessing the dataset as follows:

- ➢ **Loading the Dataset:** We loaded the mall customer dataset, which includes attributes such as CustomerID, Genre, Age, Annual Income, and Spending Score.
- ➢ **Feature Selection:** For this analysis, we selected the relevant features, namely Annual Income and Spending Score.
- ➢ **Feature Standardization:** To ensure consistency, we standardized the selected features using the StandardScaler.

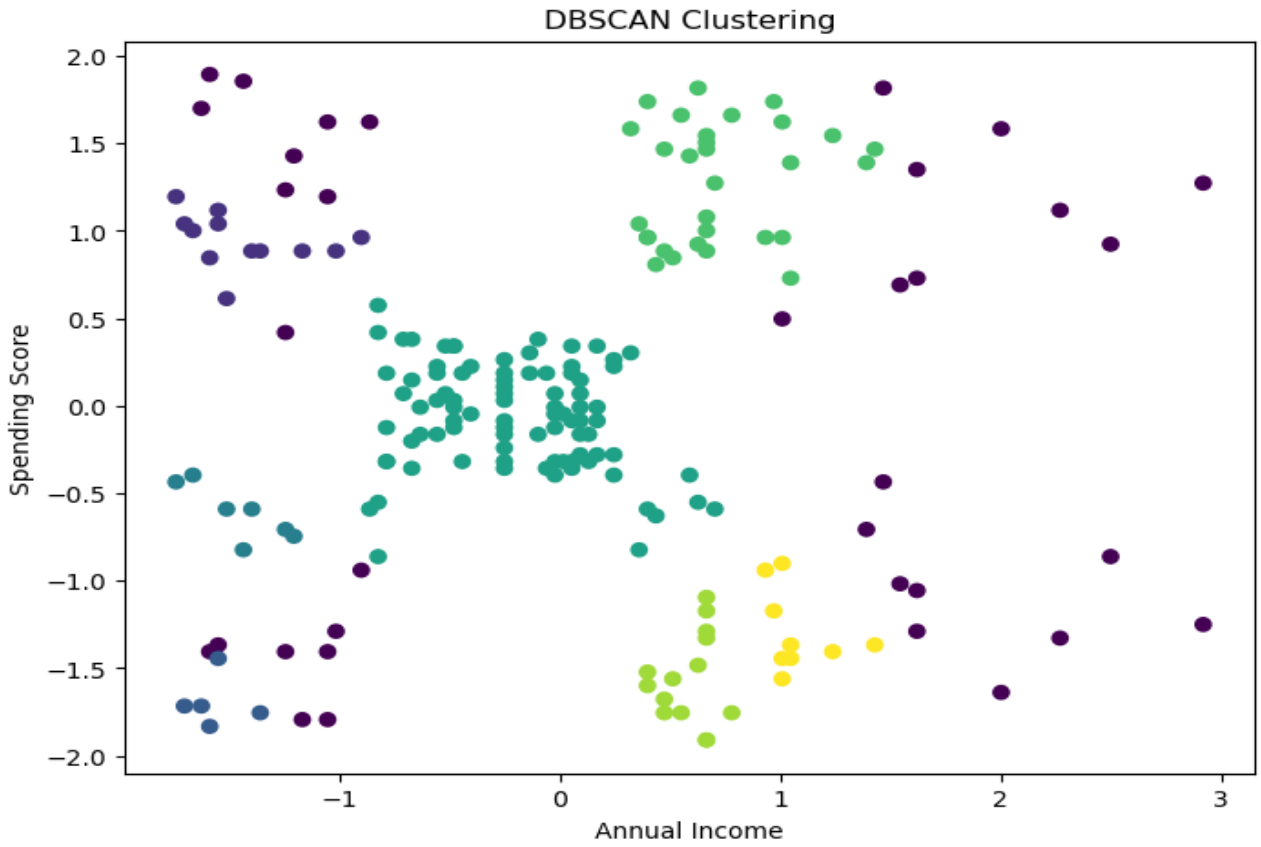
DBSCAN Clustering: DBSCAN is an unsupervised machine learning technique that identifies clusters based on the density of data points within a certain radius.

- ➢ **Objective:** To discover clusters of varying shapes and sizes, while also identifying noise points (outliers).
- ➢ **Parameters:** Key parameters include the epsilon (ε) value (radius) and the minimum number of points required within ε.

Results:

Our analysis yielded the following results:

1. We applied the DBSCAN algorithm to the dataset to identify clusters and classify data points as core points, border points, or noise points.
2. We present a scatter plot of the DBSCAN clusters, color-coding data points by their cluster assignments.


DBSCAN Clustering
Spending Score
Annual Income
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3