Bangabandhu Sheikh Mujibur Rahman Digital University

Faculty: Cyber Physical Systems

Department: IoT AND Robotics Engineering (IRE)

Course Title: Data Science

Course Code: IoT 4313

Assignment-02: Clustering

Submitted To
Nurjahan Nipa
Lecturer
Department of IRE, BDU.

Submitted By
Mohaimenul Islam Mahin
ID: 1901025
Department of IRE
Session: 2019-20

Date of Submission : 14th October 2023

Let's imagine you're owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score, which is something you assign to the customer based on your 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 you 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.

In this particular dataset we have 200 samples to study.

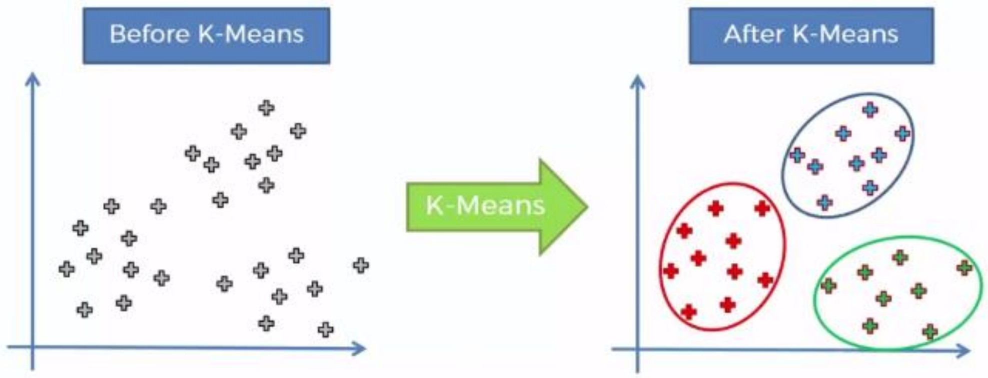
PART-(A)

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).

Introduction:

K-means clustering is an unsupervised learning technique to classify unlabeled data by grouping them by features, rather than pre-defined categories. The variable K represents the number of groups or categories created. The goal is to split the data into K different clusters and report the location of the center of mass for each cluster. Then, a new data point can be assigned a cluster (class) based on the closed center of mass. The big advantage of this approach is that human bias is taken out of the equation. Instead of having a researcher create classification groups, the machine creates its own clusters based upon empirical proof rather than assumptions.

The dataset used is clean and small for its size of course, this is a dummy. Because the dataset is clean enough, the analysis can be completed quickly. however, this is not the case with original scale data, which has a higher level of complexity and allows for the use of other additional methods.



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:

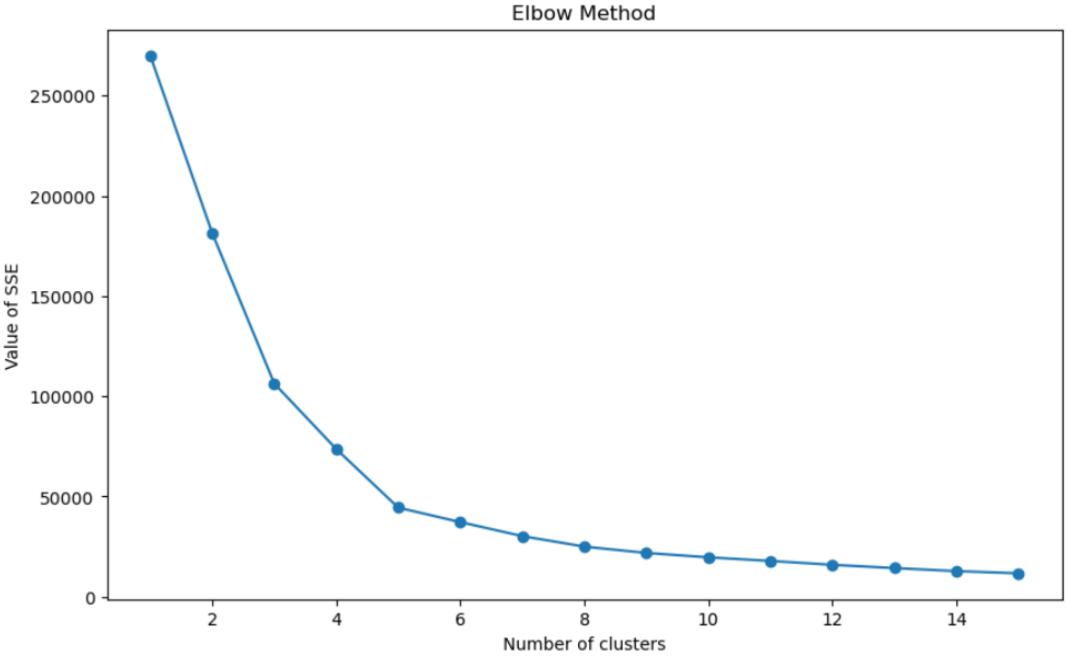
- **Range of K Values:** We considered a range of K values from 0 to 15.
- **Sum of Squared Errors (SSE):** For each K value, we calculated the corresponding SSE, which measures the distance between data points and their assigned cluster centroids.

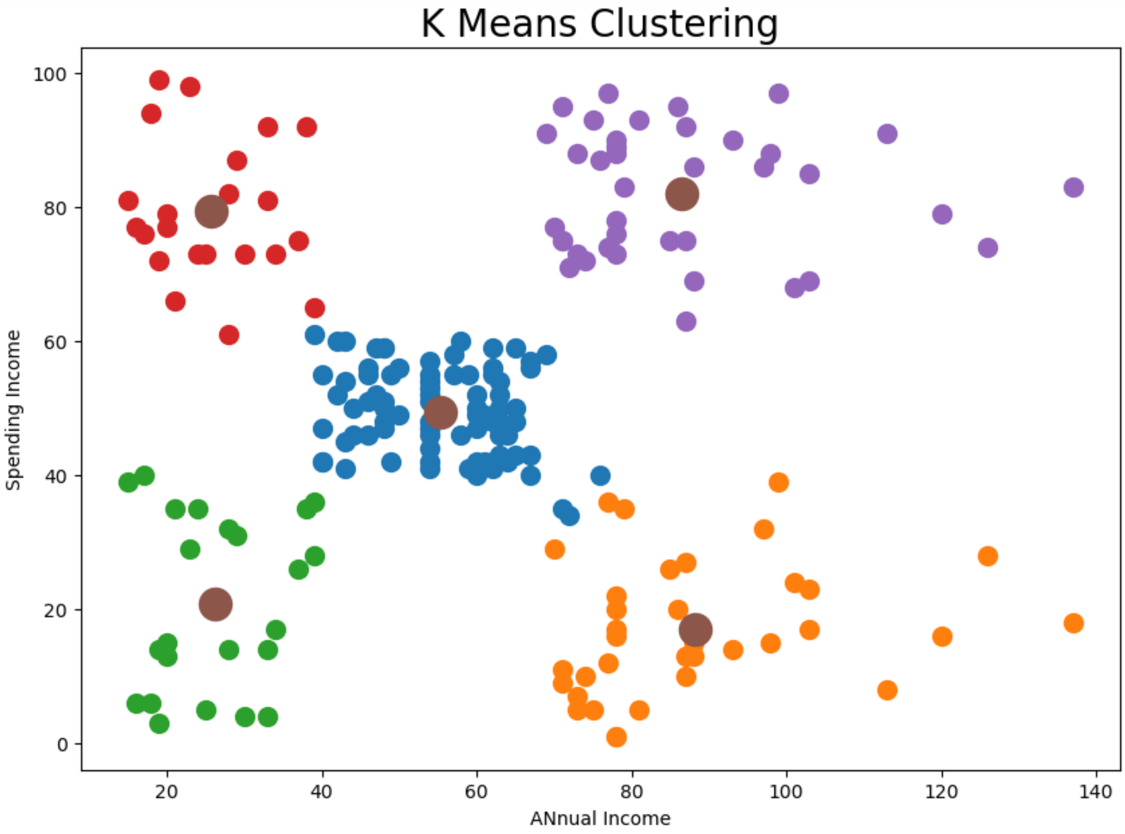
Sum of Squared Errors (SSE):

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.

Results:

We present the Elbow Method plot, depicting K values on the x-axis and SSE values on the y-axis. 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.

Introduction:

Hierarchical clustering is a unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

There are mainly two types of hierarchical clustering:

- Agglomerative hierarchical clustering
- Divisive Hierarchical clustering

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:

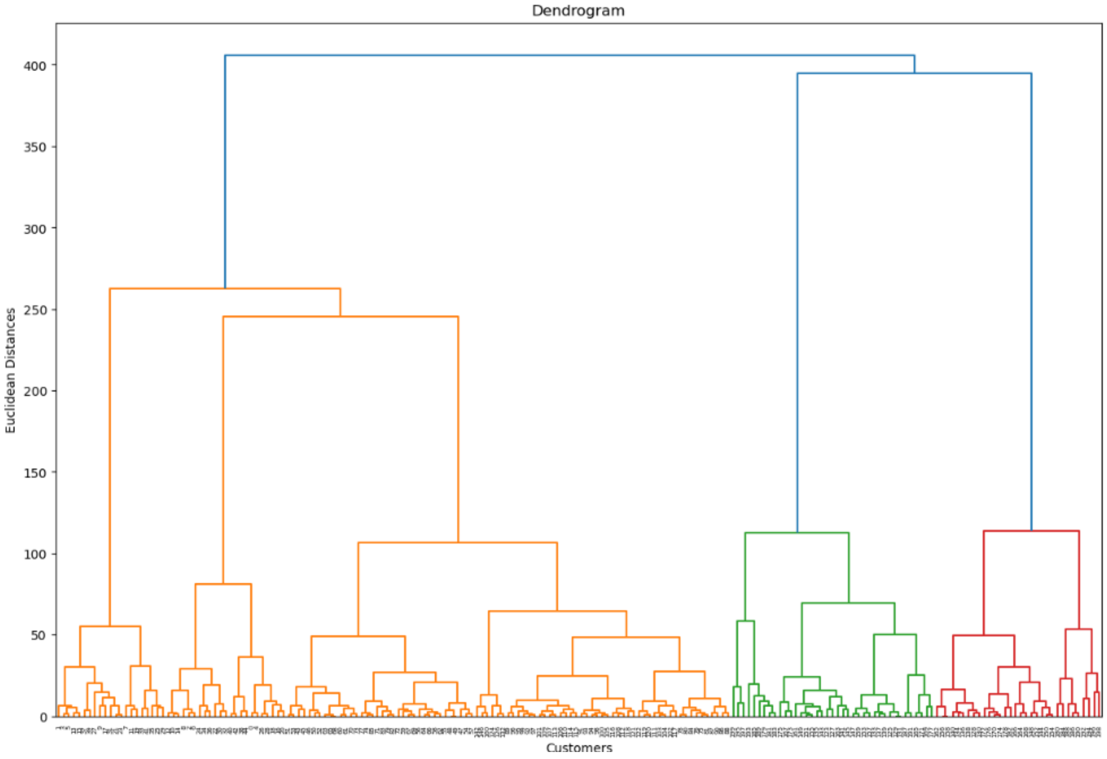
To explore the hierarchical structure of clusters, we generated dendrogram plots using the following steps:

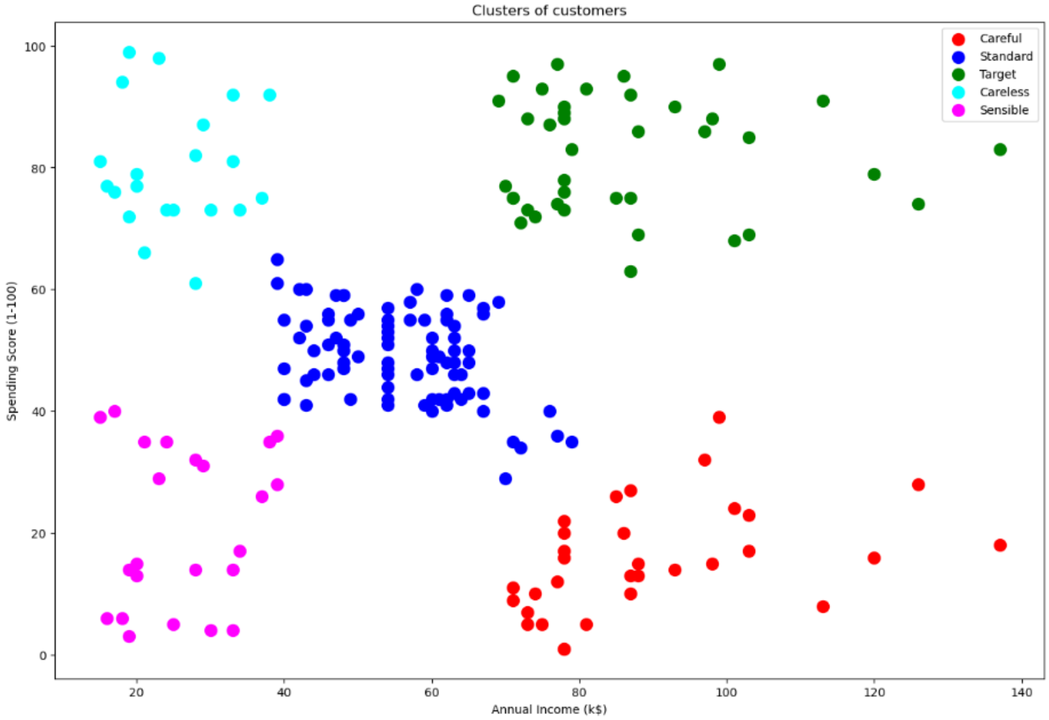
- **Linkage Method:** We used Ward's linkage method for its effectiveness in capturing cluster relationships.
- **Dendrogram Plot:** The dendrogram visualizes the hierarchy of clusters, displaying data points, branch distances, and cluster merging.

Results:

Our analysis yielded the following results:

- We present the hierarchical clustering dendrogram plot, showing the hierarchical relationships among data points.
- 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.

Introduction:

It is an unsupervised machine learning algorithm. It is used for clusters of high density. It automatically predicts the outliers and removes them. It is better than hierarchical and k-means clustering algorithm. It makes clusters based on the parameters like epsilon, min points and noise. It separately predicts the core points, border points and outliers efficiently.

DBSCAN Clustering:

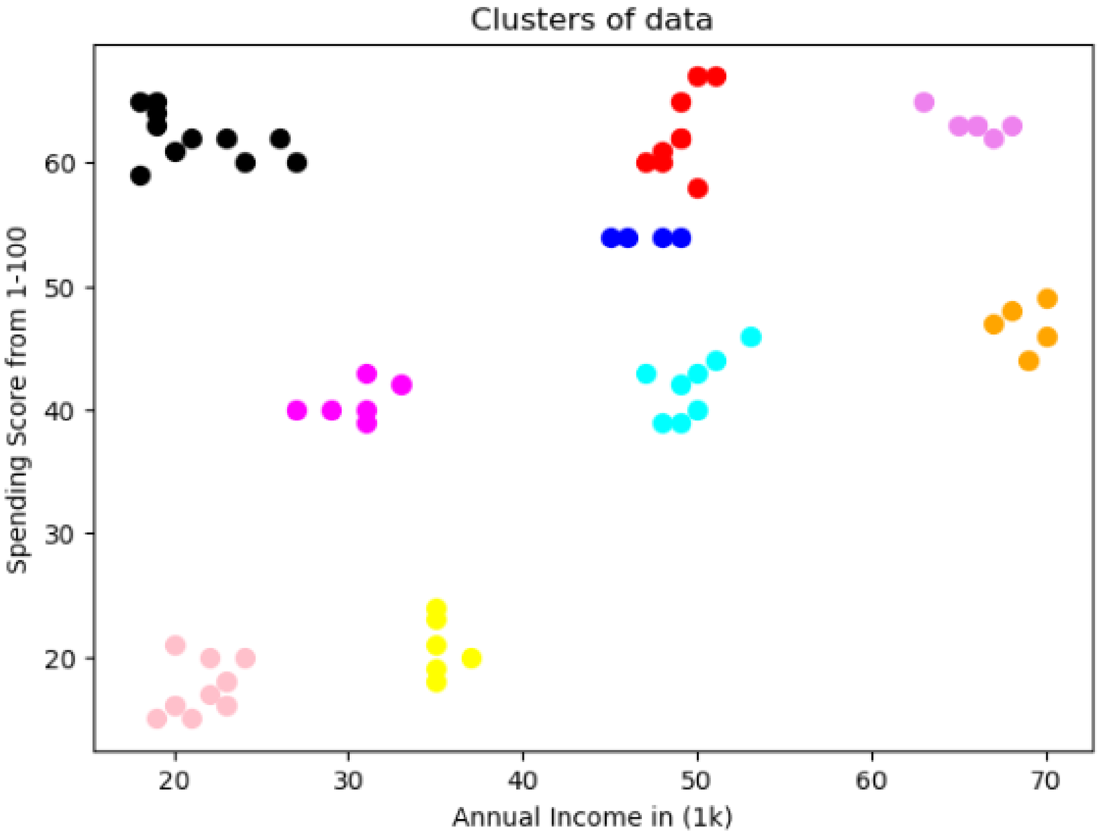
DBSCAN is an unsupervised machine learning technique that identifies clusters based on the density of data points within a certain radius.

- **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:

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



-End of the Report-