Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

Faculty of Cyber Physical System

Dept. of Internet of Things and Robotics Engineering (IRE)

Course Title: Data Science

Course Code: IOT 4313

Assignment 02: Clustering

Submitted By:
Pallab Sarkar
ID: 1801016
Session: 3rd year 2nd semester

Submitted To:
Nurjahan Nipa
Literature
Dept. Cyber Physical Systems

Date of Submission: 14-10-2023

PART (A)

K-means clustering is a simple unsupervised machine learning algorithm that divides a dataset into a predefined number of clusters. The algorithm works by iteratively assigning each data point to the cluster with the closest centroid. The centroid of a cluster is the average of all the data points in that cluster. 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).

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics.

Data Preprocessing:

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

Feature Selection: We selected the relevant features for this analysis, namely Annual Income and Spending Score.

Feature Standardization: We standardized the selected features using the StandardScaler to ensure consistency.

Advantages of Customer Segmentation:

1. Determine appropriate product pricing.
2. Develop customized marketing campaigns.
3. Design an optimal distribution strategy.
4. Choose specific product features for deployment.
5. Prioritize new product development efforts.

K Means Clustering Algorithm

K-means clustering algorithm.

1. Choose several clusters, K.
2. Initialize K centroids, either randomly or using a heuristic.
3. Assign each data point to the cluster with the closest centroid.

K Means Clustering where K=3

Environment and tools

1. scikit-learn
2. seaborn
3. numpy
4. pandas
5. matplotlib

The Elbow Method

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

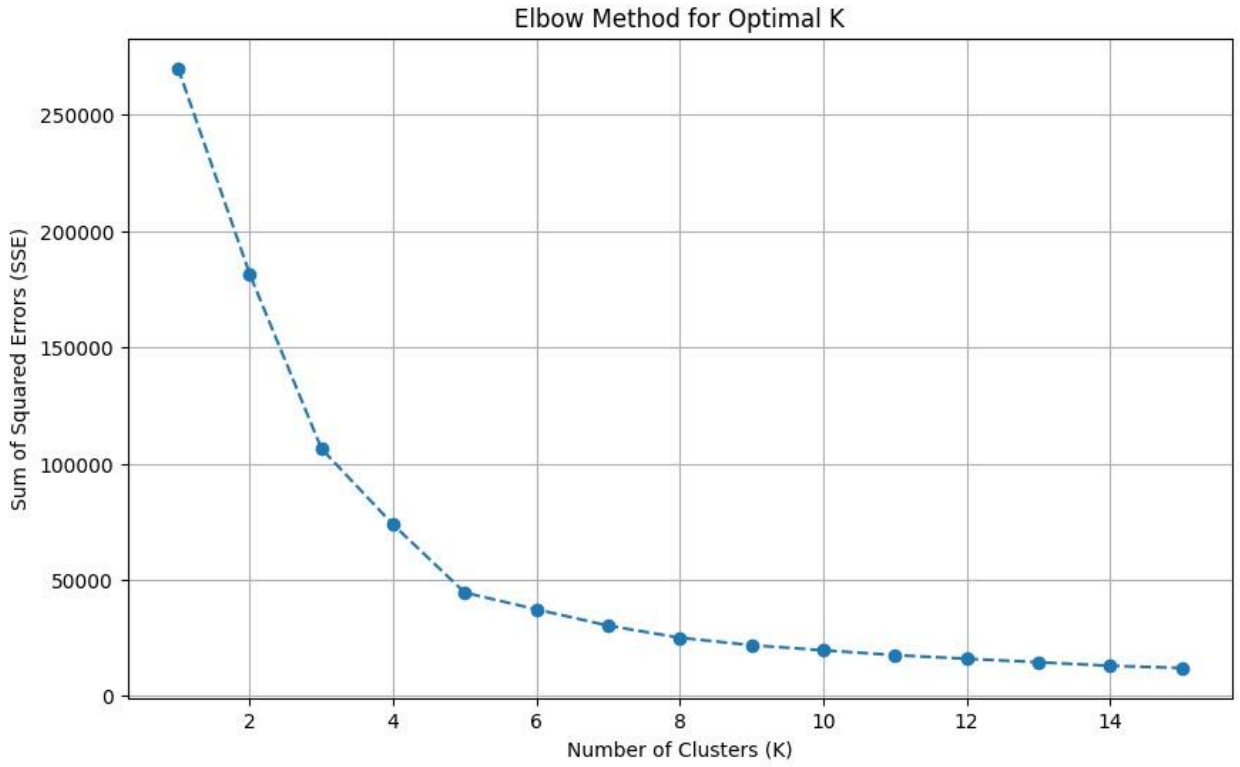
The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).

3. Plot the curve of WCSS vs the number of clusters K.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

The optimal K value is found to be 5 using the elbow method. Finally, I made a 3D plot to visualize the spending score of the customers with their annual income. The

Results



Conclusions

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The

goal of K means is to group data points into distinct non-overlapping subgroups. One of the major applications of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

PART (B)

This report presents the results of applying hierarchical clustering algorithms 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: We selected the relevant features for this analysis, namely Annual Income and Spending Score.
- Feature Standardization: We standardized the selected features using the StandardScaler to ensure consistency.

Hierarchical Clustering:

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.

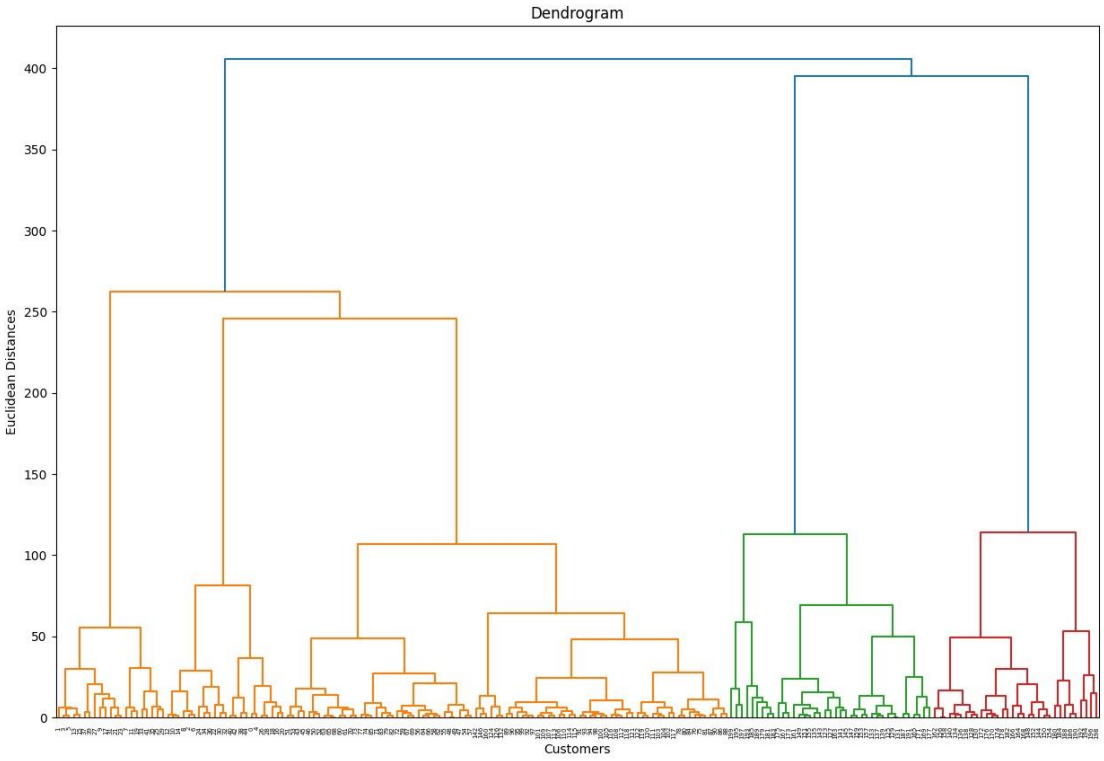
There are mainly two types of hierarchical clustering:

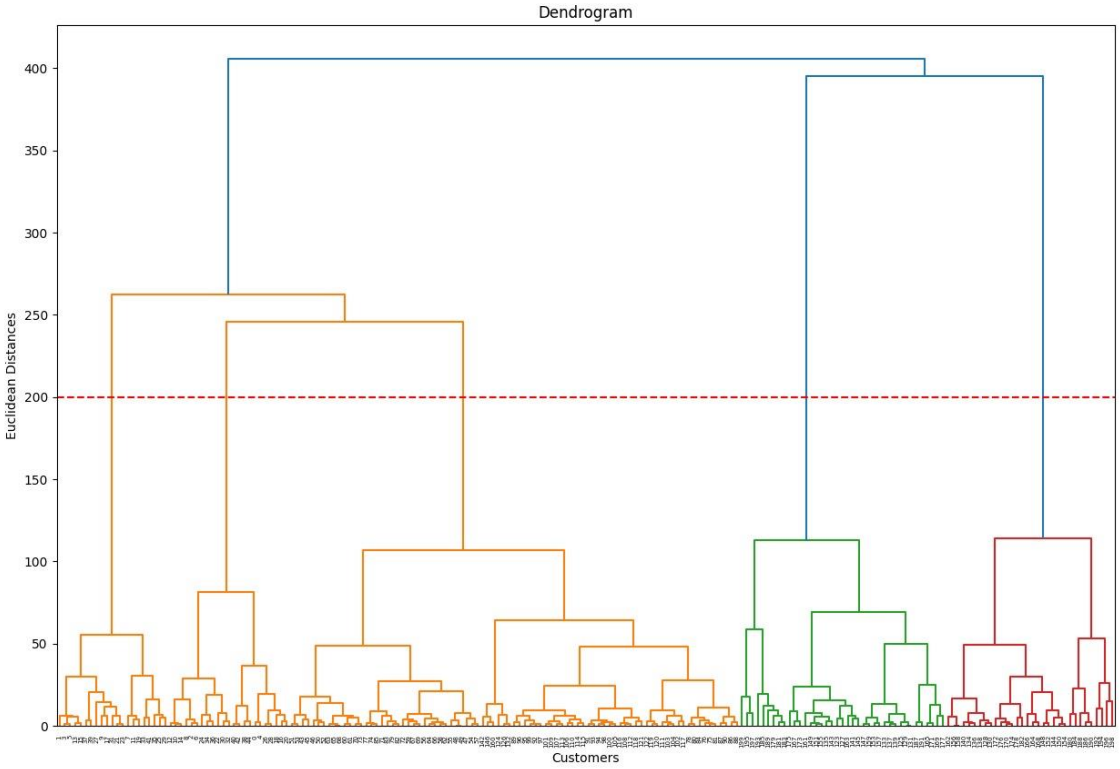
1. Agglomerative hierarchical clustering
2. Divisive Hierarchical clustering

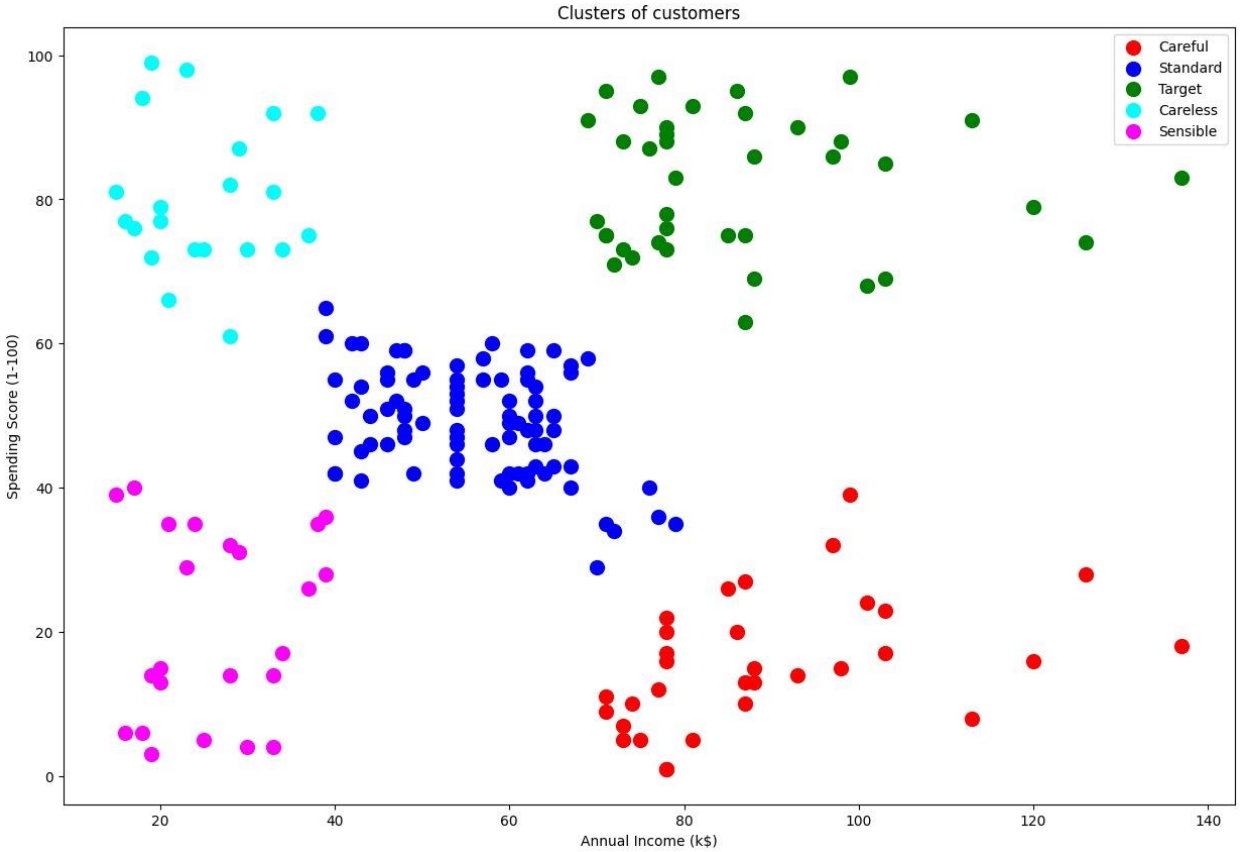
Agglomerative hierarchical clustering

It's a Bottom to Up approach clustering technique. In this initially we assign each points to be a individual clusters.

Result:




Dendrogram
Euclidean Distances
Customers
0
50
100
150
200
250
300
350
400


Clusters of customers
Careful
Standard
Target
Careless
Sensible
Spending Score (1-100)
Annual Income (k$)
0
20
40
60
80
100
20
40
60
80
100
120
140

PART (C)

Introduction

This report presents 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: We selected the relevant features for this analysis, namely Annual Income and Spending Score.
- Feature Standardization: We standardized the selected features using the StandardScaler to ensure consistency.

DBSCAN clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

DBSCAN clustering algorithm

- Choose two parameters:
- eps: the radius of the neighborhood around a point.
- min_samples: the minimum number of points within eps of a point for it to be considered a core point.
- Find all core points in the dataset.
- For each core point, find all reachable points. A reachable point is a point that is within eps of a core point or another reachable point.
- Assign all reachable points to the same cluster.
- Mark all points that are not reachable from any core point as noise.

DBSCAN clustering for customer segmentation:

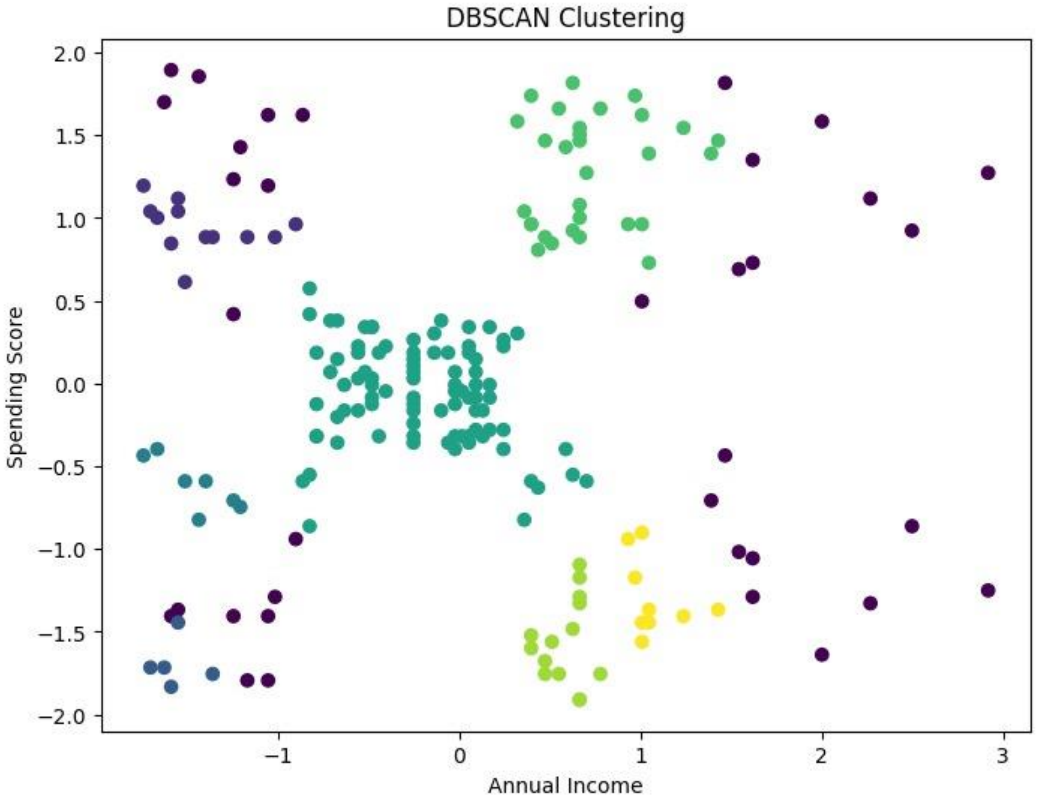
To use DBSCAN clustering for customer segmentation, we can use the following steps:

- Prepare the data. This may involve cleaning the data, removing outliers, and scaling the features.
- Choose the eps and min_samples parameters. These parameters can be tuned using a grid search or other optimization technique.
- Train a DBSCAN clustering model on the data.
- Assign each customer to a cluster based on the model's predictions.
- Analyze the clusters to understand the different customer segments.

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, colour-coding data points by their cluster assignments.



Github link: https://github.com/shasmito/dataScience-Clustring