

# FAST NATIONAL UNIVERSITY CFD CAMPUS

|  |  |
| --- | --- |
| Name: | **Mariam Fatima**  **Muskan Ghani** |
| Roll No: | **22F-3168**  **22F-3841** |
| Assignment: | **4** |
| Instructor: | **Dr. Hashim Yaseen** |
| Course: | **Machine Learning** |

# Question 1:

## Dataset Overview

The dataset includes four features (A1, A2, A3, and A4) and is used to explore three clustering techniques: K-means, K-medoids, and K-median. Each method is applied with a user-defined number of clusters K to study data grouping based on similarity.

## Methodology and Approach

### K-means Clustering

K-means clusters data by minimizing variance within clusters. It iteratively updates randomly initialized centroids to minimize squared distances until convergence.

**Advantages:**

* Fast and efficient for large datasets.
* Provides well-defined, round clusters due to its reliance on Euclidean distance.

**Disadvantages:**

* Sensitive to outliers, as it minimizes squared distances.
* May not work well with non-spherical data, leading to inaccurate clustering.

### K-medoids Clustering

K-medoids minimizes absolute deviations, using actual data points as medoids, making it more outlier-resistant.

**Advantages:**

* Less sensitive to outliers than K-means.
* Suitable for arbitrary distance measures beyond Euclidean.

**Disadvantages:**

* Computationally intensive for large datasets.
* May produce less defined clusters for complex structures.

### K-median Clustering

K-median minimizes absolute differences, offering resilience to outliers and asymmetry.

**Advantages:**

* Handles outliers better than K-means.
* Suitable for non-Euclidean distance metrics.

**Disadvantages:**

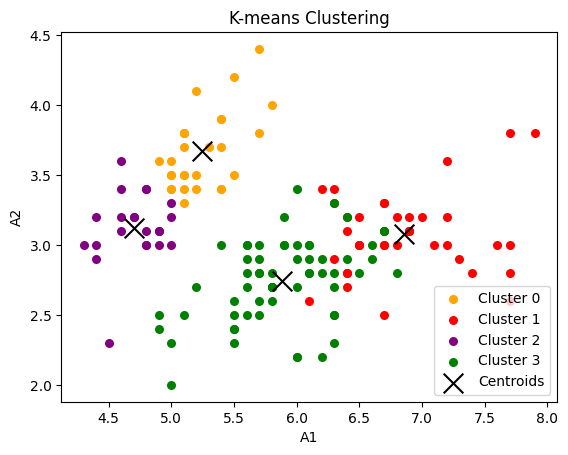
* Less efficient than K-means on very large datasets.
* Sensitive to initial cluster centers, which may lead to local minima.

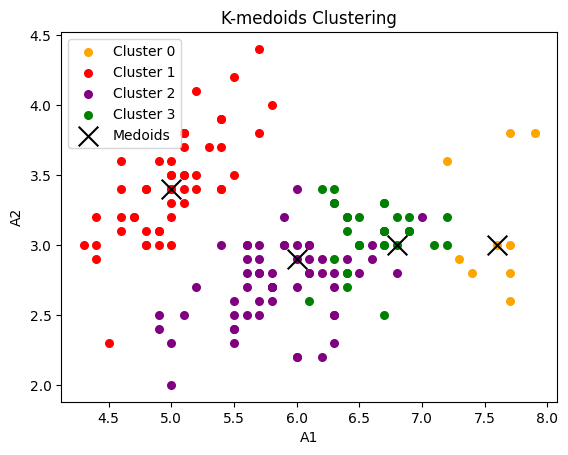
## Cluster Visualization

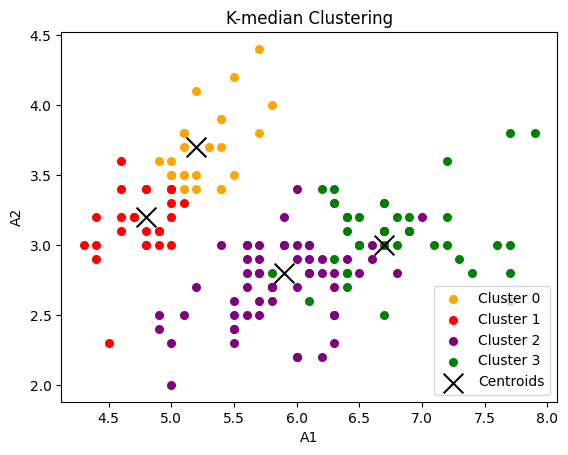
Scatter plots were generated for each method, with data points colored by their assigned clusters.

### Key Observations:

* + Cluster Shapes: K-means produced circular clusters; K-medoids and K-median formed irregular shapes.
  + Outlier Sensitivity: K-means was affected by outliers, while K-medoids and K-median were more robust.







## Comparative Analysis of Clustering Algorithms

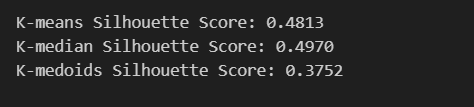
|  |  |  |  |
| --- | --- | --- | --- |
| Feature | K-means | K-medoids | K-median |
| Speed | Fast | Slower | Moderate |
| Outlier Sensitivity | High | Low | Moderate |
| Best For | Spherical clusters | Irregular shapes, outliers | Mixed shapes |

## Silhouette Score Analysis (K=4)

Silhouette scores, ranging from -1 to 1, measure cluster distinctness:

* K-median (0.4970): Highest score, indicating well-separated clusters; best fit for the dataset’s asymmetry.
* K-means (0.4813): Effective clustering, close to K-median; aligns well with spherical structures.
* K-medoids (0.3752): Lowest score; less distinct clusters, potentially less suitable for the dataset.

K-median and K-means performed best, with K-median slightly leading. K-medoids showed lower cluster clarity.



## Conclusion

The analysis highlights each method's strengths and limitations:

* K-means: Efficient for spherical clusters with minimal outliers.
* K-medoids: Better with outliers, ideal for complex shapes but slower.
* K-median: Resilient to asymmetry and outliers, though slower.

For datasets with outliers or irregular clusters, K-medoids or K-median is preferred. If speed is essential, K-means can be effective but may require careful application for non-spherical data.

# Question 2

## Hierarchical Clustering Analysis Report

## Objective:

The goal of this analysis is to apply hierarchical clustering to a dataset and visualize the results using dendrograms and scatter plots with circles representing cluster boundaries.

## Dataset:

The data, loaded from an Excel file (Data.xlsx), contains four columns (A1, A2, A3, A4). For clustering, we focus on the first two columns (A1 and A2).

## Methodology:

### Hierarchical Clustering:

We used three linkage methods to cluster the data.

### Single Linkage:

Merges clusters based on the closest points.

### Complete Linkage:

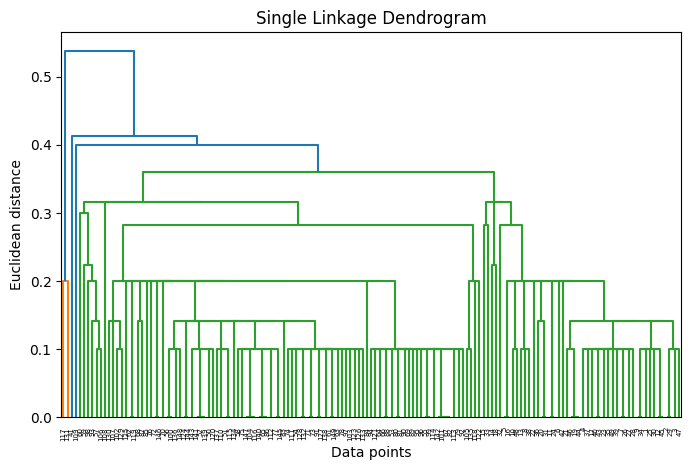
Merges clusters based on the furthest points.

### Average Linkage:

Merges clusters based on the average distance between points.

**Dendrogram Visualization**:

A dendrogram for each linkage method was plotted to show how the clusters are formed and merged. The y-axis represents the Euclidean distance between clusters.



A diagram of a linkage

Description automatically generated

A graph of a diagram

Description automatically generated with medium confidence

**Venn Diagram Visualization**:

We then used the cluster function to define clusters based on a distance threshold. For each cluster, we plotted the data points with unique colors and added circles around each cluster's points to highlight their boundaries.

**Circles**:

The radius of each circle was calculated to cover all points in the cluster, helping to visually define the cluster's area.

A graph of a diagram

Description automatically generated with medium confidence

A diagram of a cluster

Description automatically generated

A graph showing multiple colored circles

Description automatically generated

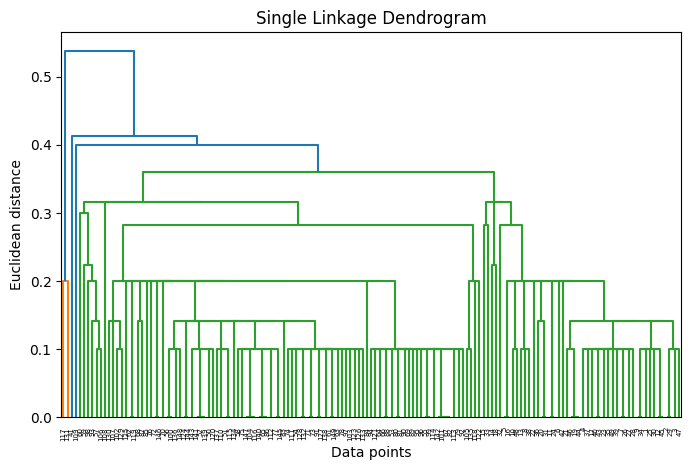
## Results:

**Dendrograms**:

Each linkage method shows different cluster structures:

**Single Linkage**:

Produces more elongated, stringy clusters.



**Complete Linkage**:

Creates compact, well-defined clusters.

A graph of a linkage

Description automatically generated

**Average Linkage**:

Balances between the two, offering a mix of compact and spread-out clusters.

A graph of a diagram

Description automatically generated with medium confidence

**Cluster Boundaries**:

For each linkage method, we visualized the clusters with colored points and boundary circles. This makes it easier to understand how the data points are grouped together.

**Conclusion**:

Hierarchical clustering reveals how data points group together based on different criteria. The visualizations provide a clear understanding of cluster boundaries, making it easier to interpret the data. Adjusting the distance threshold allows us to explore the data at different levels of granularity, revealing various patterns.