Clustering Analysis on the Iris Dataset

R.Pridhurshika (S/19/484)

1 INTRODUCTION

Clustering is an unsupervised learning technique that groups similar data points into clusters based on their similarity or distance. It plays a crucial role in pattern discovery, data compression, and exploratory analysis. In this study, we apply **K-Means** and **Hierarchical Clustering** to the well-known Iris dataset, using various distance metrics and linkage criteria, to compare their performance and recommend the most suitable approach.

1.1 Objective

Explore clustering techniques on the Iris dataset using K-Means and Hierarchical Clustering with different distance metrics and linkage criteria.
Compare results and propose the best clustering model

1.2 Data Description

The Iris dataset contains:

- 150 samples from three species of Iris flowers (*setosa*, *versicolor*, *virginica*).
- 4 numerical features: sepal length, sepal width, petal length, petal width (in cm).
- Balanced dataset with 50 samples per class.

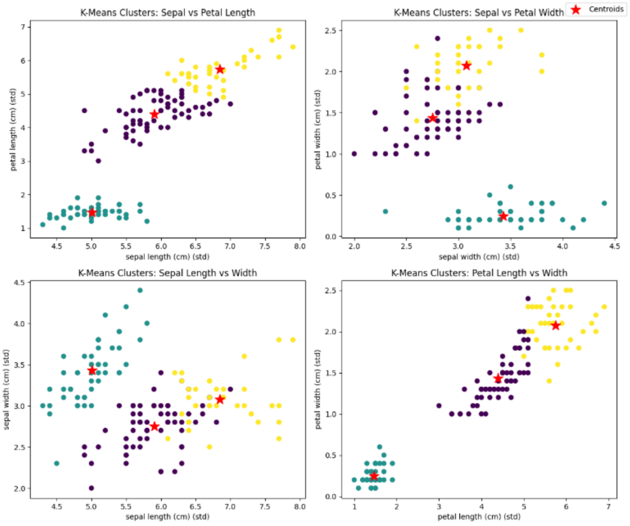
All features were standardized to have zero mean and unit variance.

2 METHODOLOGY

1. **Preprocessing:** Standardized all features using z-score normalization.
2. **Clustering Techniques:**
   - K-Means with Euclidean distance, $k = 3$.
   - Hierarchical clustering with:
     - Distance metrics: Euclidean, Cityblock (Manhattan), Minkowski.
     - Linkage criteria: Single, Complete, Average.
3. **Evaluation:** Used Silhouette Score to assess cluster separation (range: -1 to 1).

3 RESULTS AND DISCUSSION

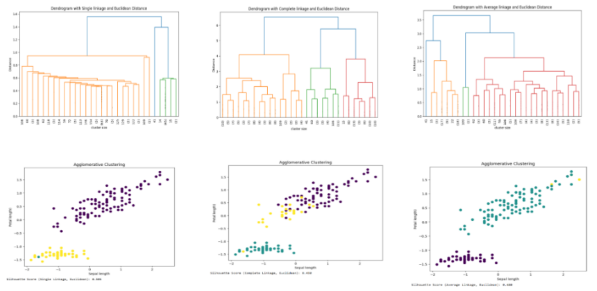
3.1 K-Mean Clustering



The clustering achieved a **silhouette score of 0.553**, which indicates a **moderately good clustering quality**. This score suggests that the clusters are reasonably well separated, but not perfectly distinct. Some overlap is visible, especially between the purple and yellow clusters. Overall, the model captures meaningful structure in the data, but cluster boundaries are not entirely sharp.

3.2 Hierarchical Clustering

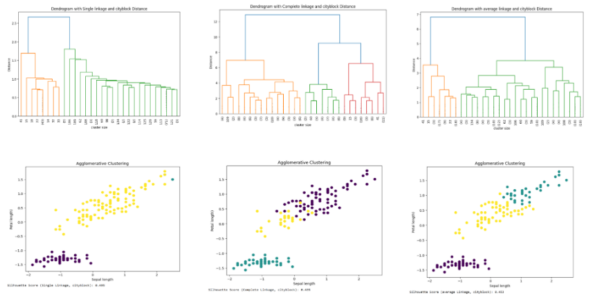
1) Euclidean Distance



a) Single Linkage b) Complete Linkage c) Average Linkage

Figure 1: Hierarchical Clustering with Euclidean distance

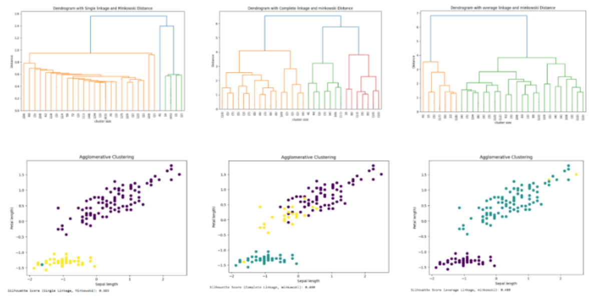
2) Manhattan (City Block) Distance



a) Single Linkage b) Complete Linkage c) Average Linkage

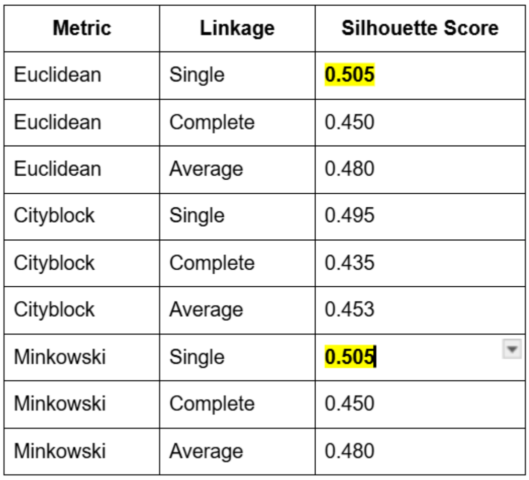
Figure 2: Hierarchical Clustering with Manhattan distance

3) Minkowski Distance



a) Single Linkage b) Complete Linkage c) Average Linkage

Figure 3: Hierarchical Clustering with Minkowski distance



Metric	Linkage	Silhouette Score
Euclidean	Single	0.505
Euclidean	Complete	0.450
Euclidean	Average	0.480
Cityblock	Single	0.495
Cityblock	Complete	0.435
Cityblock	Average	0.453
Minkowski	Single	0.505
Minkowski	Complete	0.450
Minkowski	Average	0.480

4 CONCLUSIONS

Both K-Means and Hierarchical Clustering showed moderate clustering quality on the Iris dataset (silhouette scores 0.4–0.6). K-Means produced slightly more compact clusters, while hierarchical clustering—especially with Euclidean or Minkowski single linkage—achieved higher silhouette scores and revealed nested relationships. Overall, both methods captured meaningful patterns despite imperfect cluster separation.