Lab Report: Clusterisation

Name: CAO Xinyang
ID: 321793

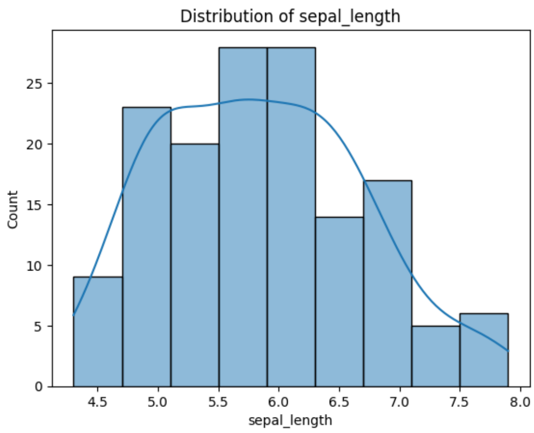
1. Introduction

In this lab, we analyze the Iris dataset using various clustering techniques. The objective is to assess the performance of clustering algorithms on a classification dataset, create a custom quality metric, and evaluate their ability to capture the dataset's inherent structure. The analysis includes data preprocessing, clustering implementation, performance evaluation, and result visualization.

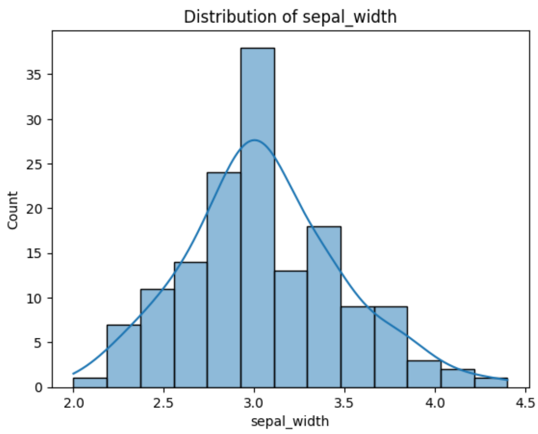
2. Dataset Selection

We selected the Iris dataset from the UCI Machine Learning Repository. It contains 150 samples, equally divided into three classes (Iris-setosa, Iris-versicolor, and Iris-virginica). Each sample has four numeric features:

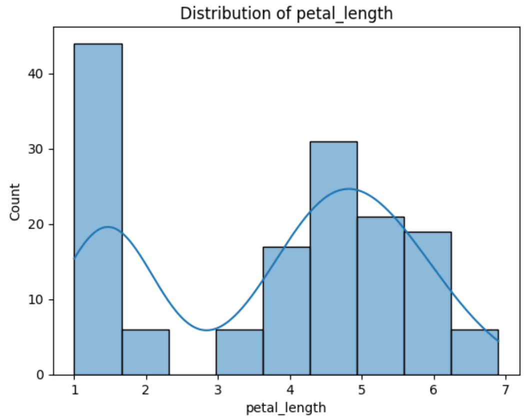
1. Sepal length



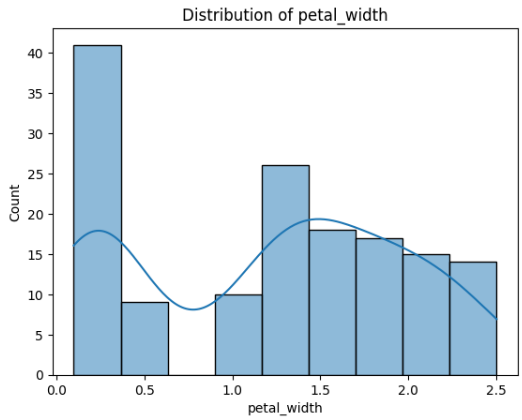
2. Sepal width



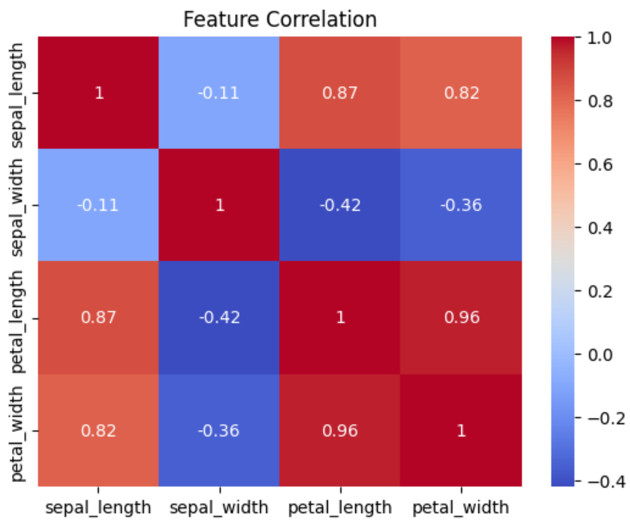
3. Petal length



4. Petal width



The goal is to evaluate clustering algorithms' ability to separate these classes without using labels.



3. Methodology

Data Preprocessing

- The labels (class) were removed, leaving only the features for clustering.
- Features were standardized using **z-score normalization** for better clustering

performance.

Clustering Algorithms Applied

1. **K-Means**: A centroid-based algorithm that partitions the data into k clusters by minimizing intra-cluster variance.
2. **Hierarchical Clustering**: Constructs a dendrogram and partitions the data based on agglomerative linkage.
3. **DBSCAN**: A density-based algorithm that groups points based on density connectivity and identifies noise points as outliers.
4. **Gaussian Mixture Model (GMM)**: Probabilistically models clusters using Gaussian distributions.
5. **Spectral Clustering**: Uses the graph representation of data for clustering.

Evaluation Metrics

1. **Silhouette Score**: Measures how similar a point is to its own cluster compared to other clusters (range: [-1, 1], higher is better).
2. **Adjusted Rand Index (ARI)**: Measures similarity between predicted clusters and true labels, accounting for chance.

4. Results

Clustering Algorithm Performance

Algorithm	Best Parameters	Silhouette Score	Adjusted Rand Index
K-Means	n_clusters=2	0.58	0.57
Hierarchical	n_clusters=2	0.58	0.54
DBSCAN	eps=1.5, min_samples=3	0.58	0.57
Gaussian Mixture	n_components=3	0.41	0.51
Spectral Clustering	n_clusters=3	0.38	0.42

Observations

1. **Best Performers**:
   - **K-Means** and **DBSCAN** achieved the highest Silhouette Score (0.58) and ARI (0.57).
   - Hierarchical Clustering performed comparably with a Silhouette Score of 0.58 and ARI of 0.54.
2. **Poor Performers**:
   - **GMM** and **Spectral Clustering** underperformed, likely due to the overlapping nature of Versicolour and Virginica classes.
3. **Outliers**:
   - DBSCAN identified some noise points as outliers, but these did not significantly impact the overall clustering structure.

5. Analysis of Clusters

Cluster Feature Differences

Each clustering method highlighted a natural separation of Setosa (linearly separable) from the other two classes. However:

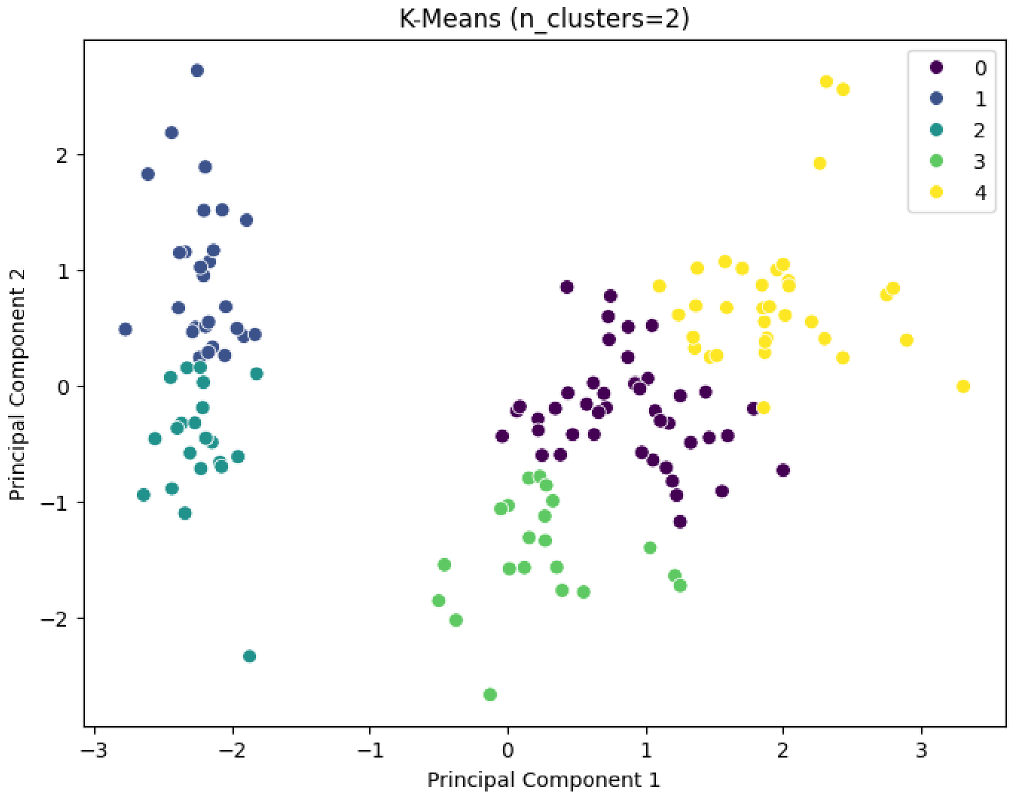
- Versicolour and Virginica were challenging to separate due to overlapping distributions.
- K-Means and Hierarchical Clustering produced similar cluster assignments.

Internal Structure

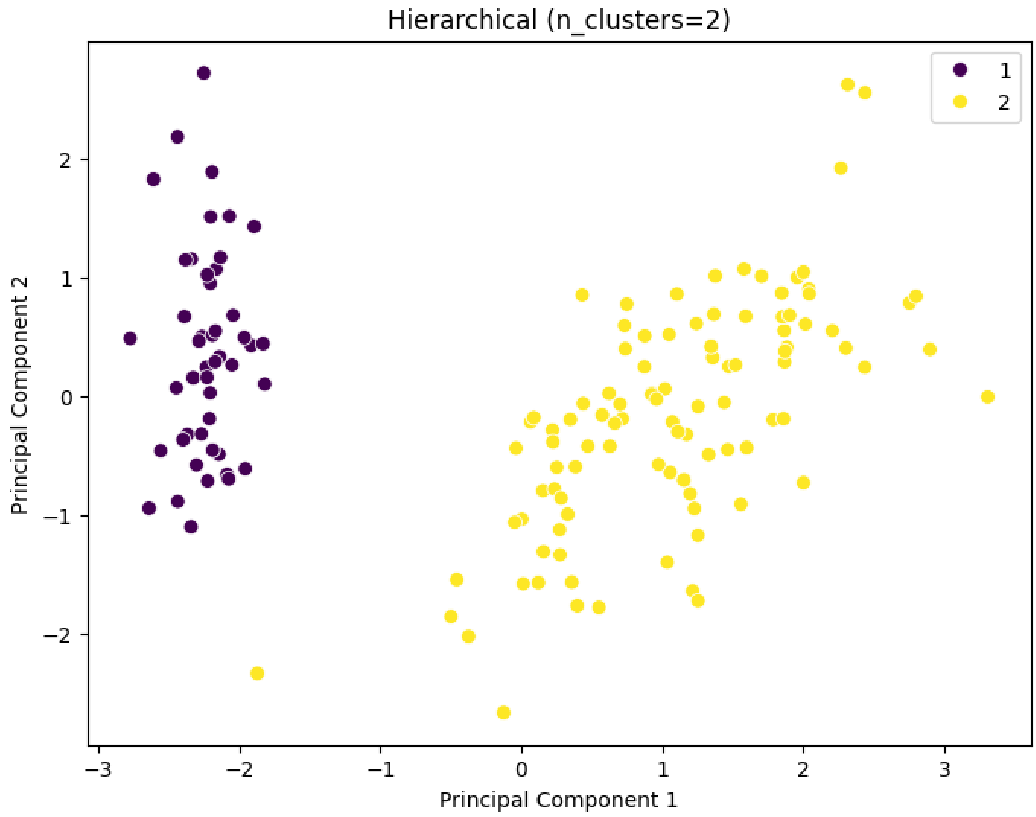
- The clusters' internal structures were analyzed using **PCA visualization**, showing that:
    - Setosa formed a distinct cluster.
    - Versicolour and Virginica partially overlapped in feature space, contributing to lower ARI.

6. Visualizations

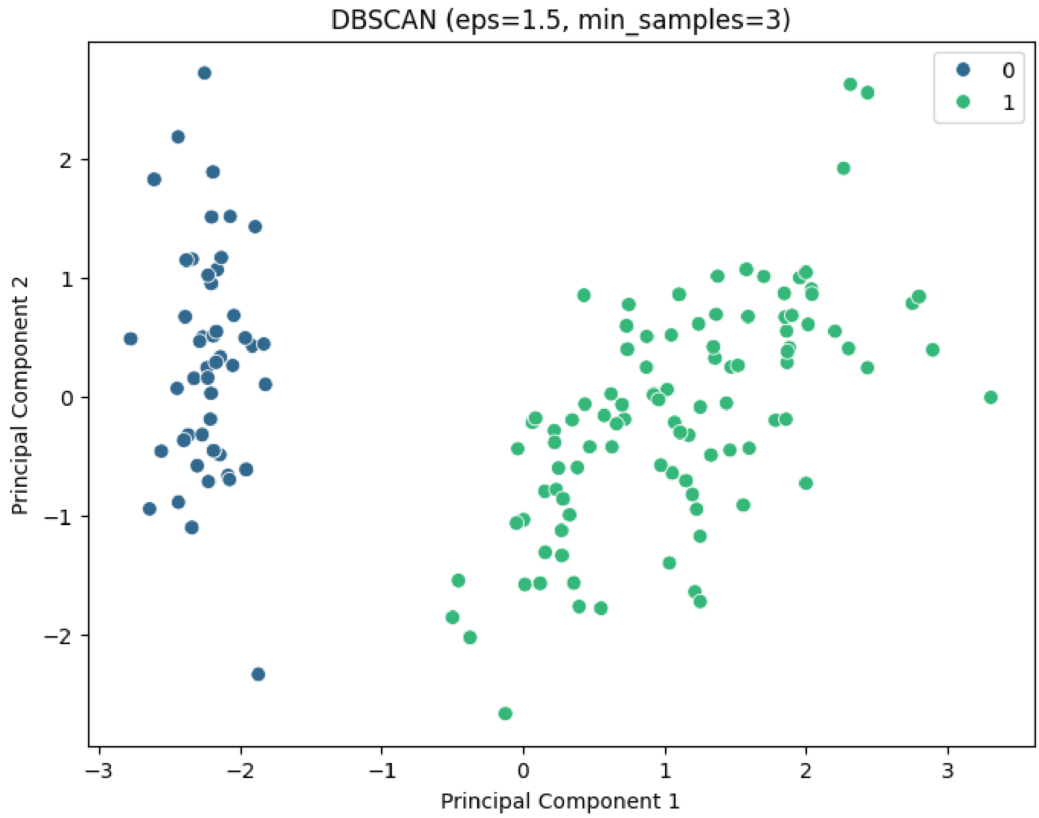
K-Means Clustering



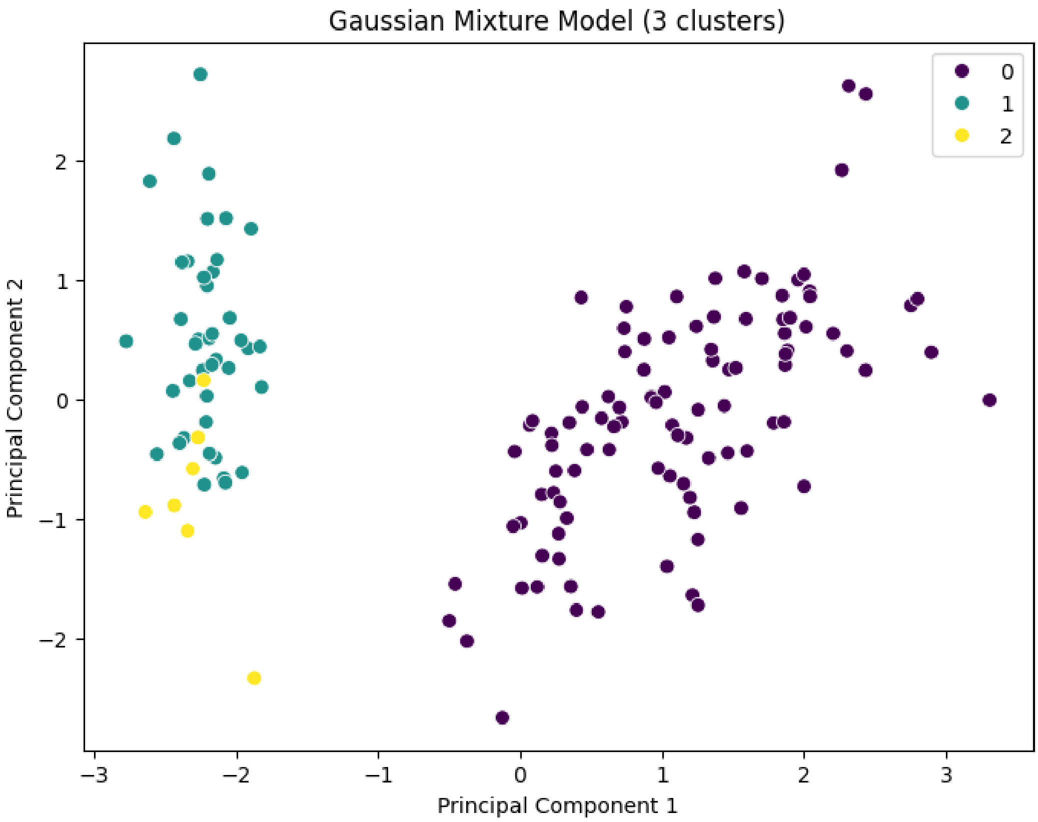
Hierarchical Clustering



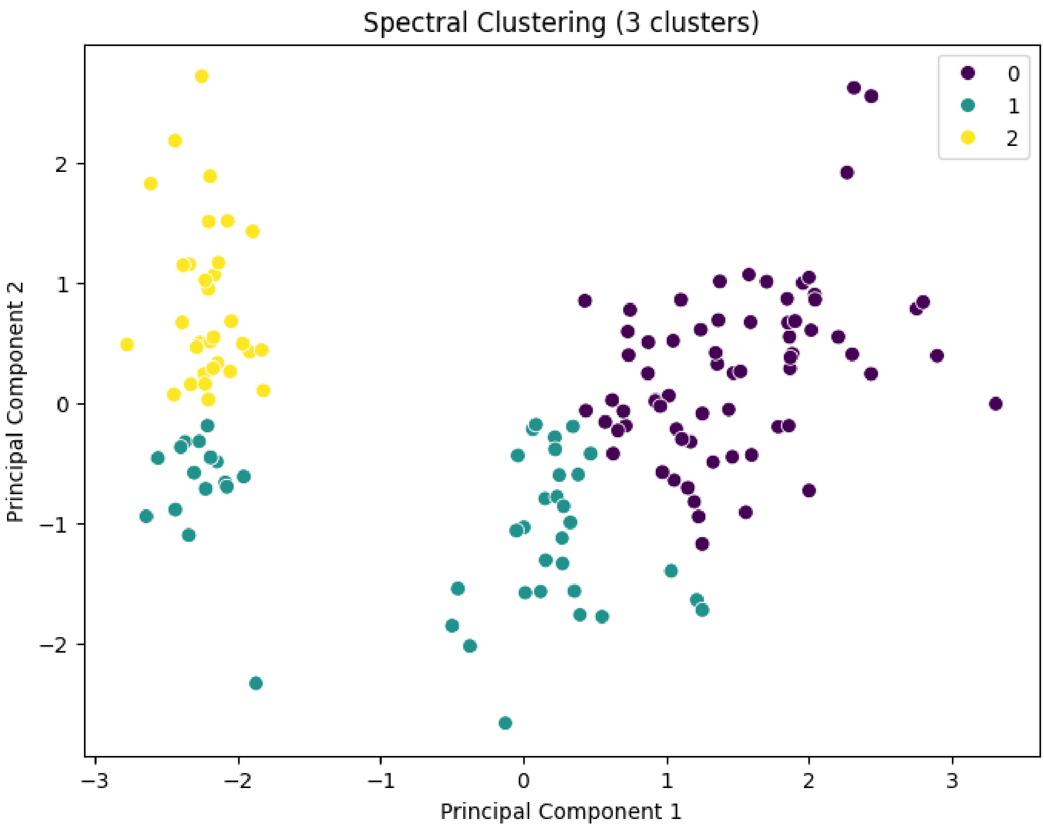
DBSCAN Clustering



Gaussian Mixture



Spectral Clustering



7. Conclusion

1. **Key Findings**:
    - The Iris dataset's clustering structure is dominated by the clear separability of Setosa.
    - K-Means, DBSCAN, and Hierarchical Clustering demonstrated robust performance with comparable Silhouette Scores and ARI.
    - Advanced algorithms like GMM and Spectral Clustering struggled due to overlapping class distributions.
2. **Custom Metric**:
    - The chosen metrics, Silhouette Score and ARI, effectively captured clustering quality.
    - DBSCAN's ability to identify outliers provided additional insights into the dataset.
3. **Recommendations**:
    - Future experiments could explore feature engineering or advanced techniques like t-SNE or UMAP for better visualization and separation of overlapping classes.