A5. Clustering

## Clustering for the Diabetes Dataset

## Partition Clustering

**K-Means Clustering**: This algorithm partitions data into ***k*** clusters. Each data point belongs to the cluster with the nearest mean, which serves as the cluster center.

**Initialization**: Define the number of clusters kk.

**Assignment Step**: Assign each data point to the nearest cluster center.

**Update Step**: Calculate the new mean of each cluster.

**Convergence**: Repeat the assignment and update steps until the cluster centers do not change significantly.

### Evaluation Metrics:

**Silhouette Score**:

Measures how similar a data point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.

**Adjusted Rand Index (ARI)**:

Measures the similarity between the true labels and the clustering results. A higher score indicates better clustering performance.

**K-Medoids Clustering**:

Similar to K-means but uses medoids (actual data points) instead of means to define clusters.

**Initialization**:

Select kk medoids.

Assignment Step: Assign each data point to the nearest medoid.

Update Step: Select new medoids that minimize the total distance within the cluster.

Evaluation Metrics: Same as K-means (Silhouette Score and ARI).

## Hierarchical Clustering

Agglomerative Hierarchical Clustering: Builds a hierarchy of clusters by iteratively merging the closest pairs of clusters.

Linkage Methods: Determines how the distance between clusters is calculated (e.g., single, complete, average).

Dendrogram: A tree-like diagram showing the arrangement of clusters produced by hierarchical clustering.

Evaluation Metrics: Silhouette Score for different numbers of clusters.

**Density-Based Clustering**

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Identifies clusters based on the density of data points.

Parameters:

* ***eps***: The maximum distance between two samples for them to be considered as in the same neighborhood.
* ***min\_samples***: The number of samples in a neighborhood for a point to be considered a core point.

Evaluation Metrics: Number of clusters, Silhouette Score, and ARI.

## Detailed Data Analysis

**K-Means Clustering**

***Silhouette Score***: 0.1795

* This indicates that the clusters are not well-separated; points in the same cluster are not significantly more similar to each other than to points in other clusters.

***ARI Score***: 0.1219

* This indicates a low similarity between the true labels and the clustering results, suggesting that K-means is not capturing the natural structure of the data effectively.

**K-Medoids Clustering**

***Silhouette Score***: 0.1212

* This is lower than the K-means score, suggesting that the clusters are even less well-defined.

***ARI Score***: 0.0804

* This indicates an even lower agreement between the true labels and the clustering results compared to K-means.

**Hierarchical Clustering**

Silhouette Scores for Different Numbers of Clusters:

* 2 Clusters: 0.1568
* 3 Clusters: 0.1706
* 4 Clusters: 0.1695
* 5 Clusters: 0.1778
* 6 Clusters: 0.1345
* 7 Clusters: 0.1481

These scores suggest that 3 or 5 clusters might provide the best separation of data points, although the scores are generally low.

**DBSCAN Clustering**

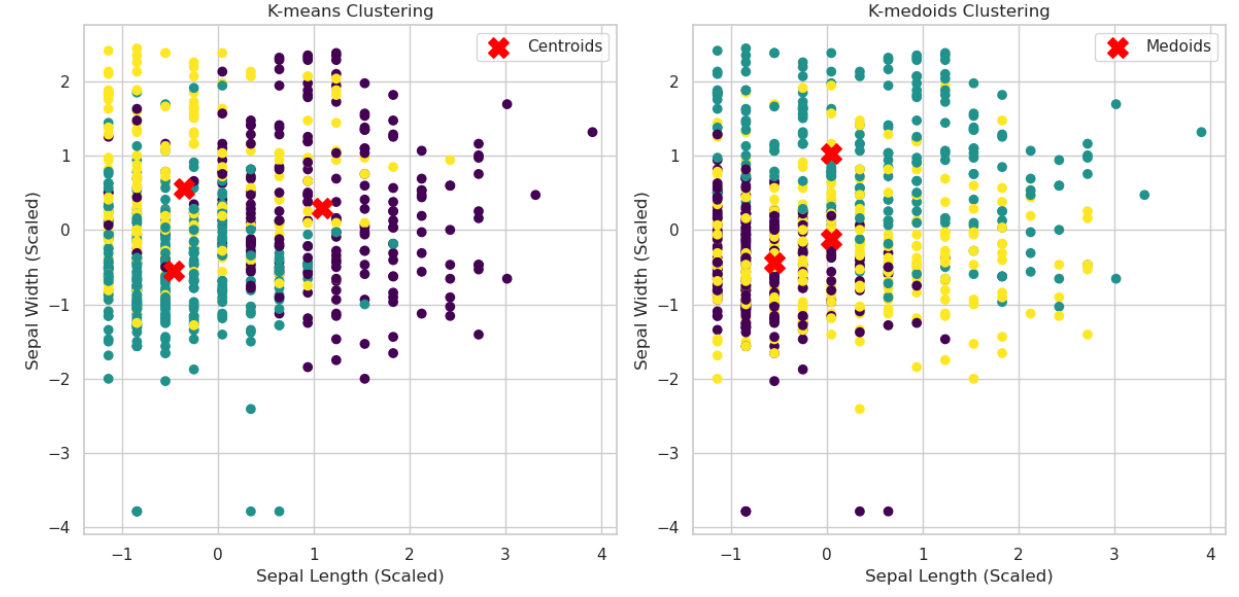
Different Combinations of eps and min\_samples:

* Most combinations resulted in 0 clusters, which is indicative of DBSCAN's sensitivity to parameter selection.
* A few combinations, such as (eps=0.5, min\_samples=2), identified more clusters, but the Silhouette Scores were negative or close to 0, indicating poor cluster quality.

## Interpretation of Results

**K-Means and K-Medoids**

* The low Silhouette and ARI scores indicate that these partitioning methods are not well-suited for this dataset.
* Clusters identified do not correspond well with the natural groupings in the data, making it hard to derive meaningful insights.



**Hierarchical Clustering**

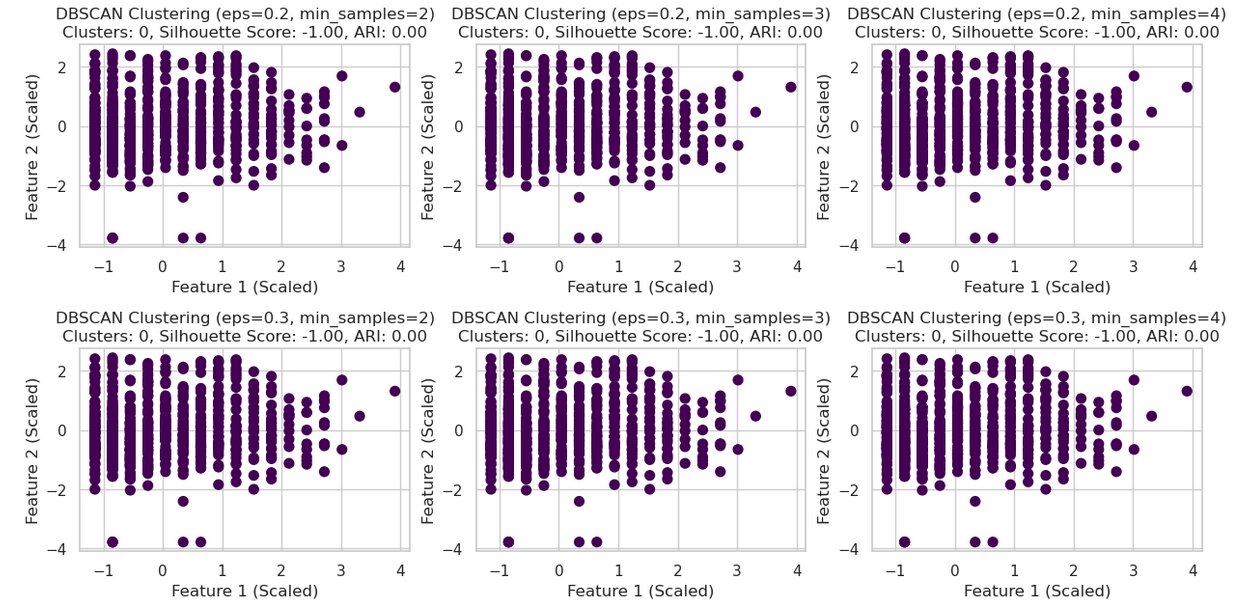
Hierarchical clustering with 3 or 5 clusters provided slightly better silhouette scores, suggesting a somewhat better-defined structure.

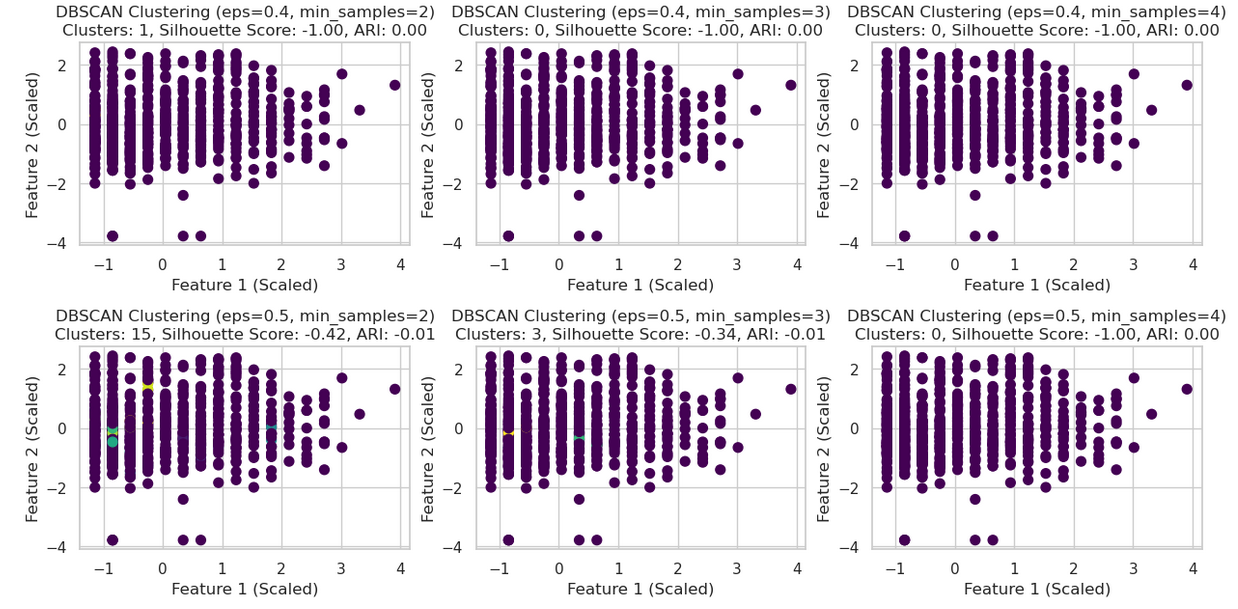
The dendrograms can provide insights into the hierarchical relationships between data points, potentially useful for understanding subgroups within the data.

**DBSCAN Clustering**

DBSCAN's performance indicates that the data does not have well-defined density-based clusters.

The choice of eps and min\_samples heavily influences the outcome, and in most cases, DBSCAN failed to identify meaningful clusters.





## Choosed Methods

**StandardScaler**:

Scaling was necessary to ensure that features contribute equally to the distance calculations used in clustering algorithms.

**Evaluation Metrics**: Silhouette Score and ARI were chosen for their ability to measure cluster cohesion and the agreement with true labels, respectively.

**DBSCAN Parameter Tuning**:

A range of eps and min\_samples values were tested to explore the sensitivity of the algorithm to these parameters and identify the best possible clustering.

## Results interpretation:

The clustering results suggest that the data does not naturally group into well-defined clusters using these methods.

For meaningful insights, more sophisticated techniques or domain-specific feature engineering might be necessary.