DMG ASSIGNMENT - 2 CLUSTERING

Jatin Tyagi	2020381
Pratyush Jain	2020369
Ritvik Pendyala	2020096
Yatish Garg	2020162
Vatsal	2020148

CONTRIBUTIONS

Jatin Tyagi: Contributed in developing the visualisation part in this assignment and also contributed in Question 1,2,3.

Pendyala Ritvik: Contributed in developing the visualisation part in this assignment and also contributed in Question 3,5.

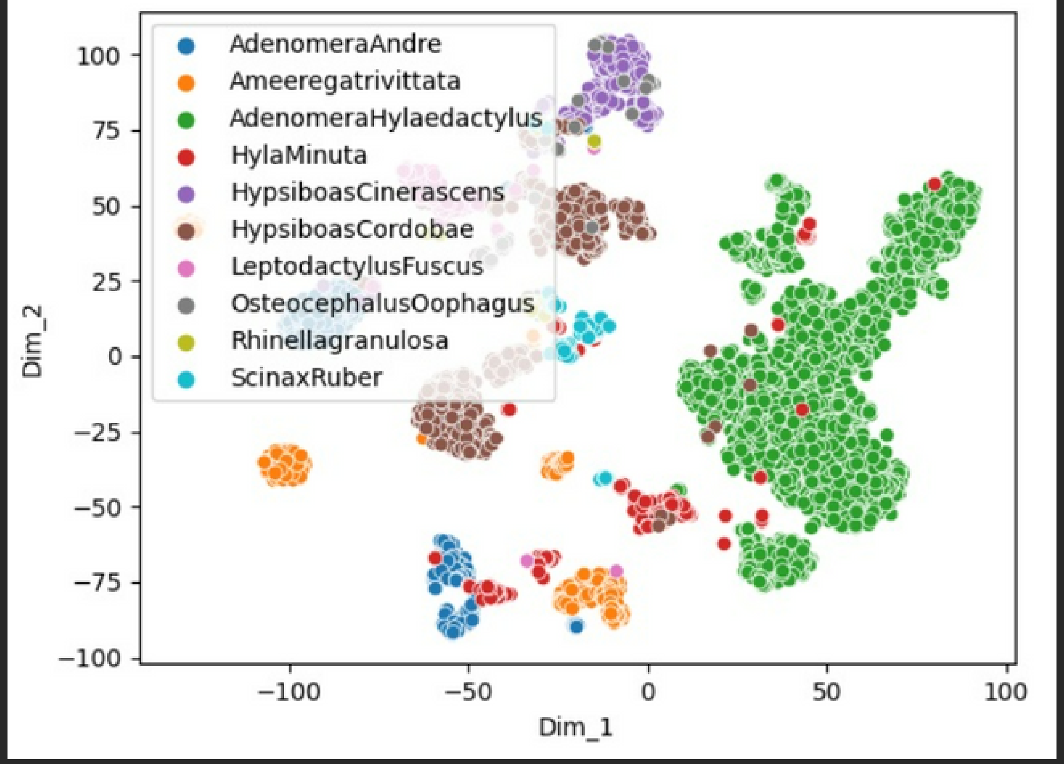
Pratyush Jain: Made the Readme file for this assignment, he helped in choosing the datasets for the assignment and helped in Question 2,3,5.

Vatsal Lakhmani: Made the PPT for this assignment and made contributions in visualisation part and Question 2,5.

Yatish Garg: Made the Report file for this assignment and also made contributions in visualisation part and Question 2,4.

DATASETS

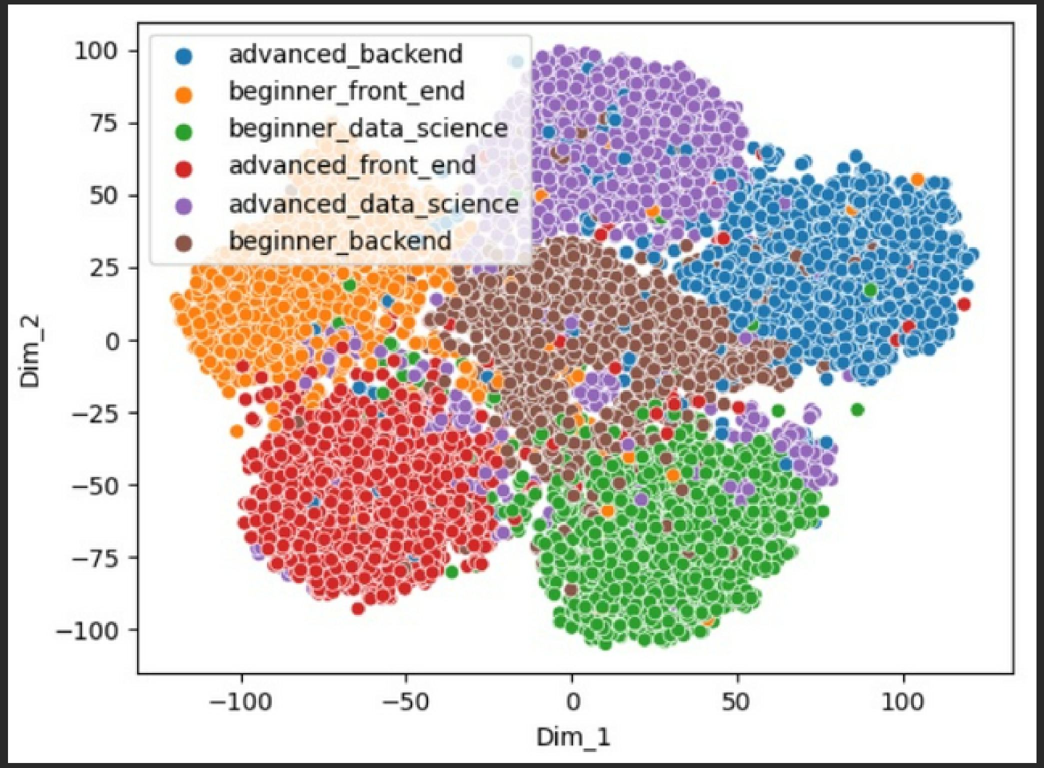
Density based clustering



Dataset 1

7195 rows
22 features
Target: Species

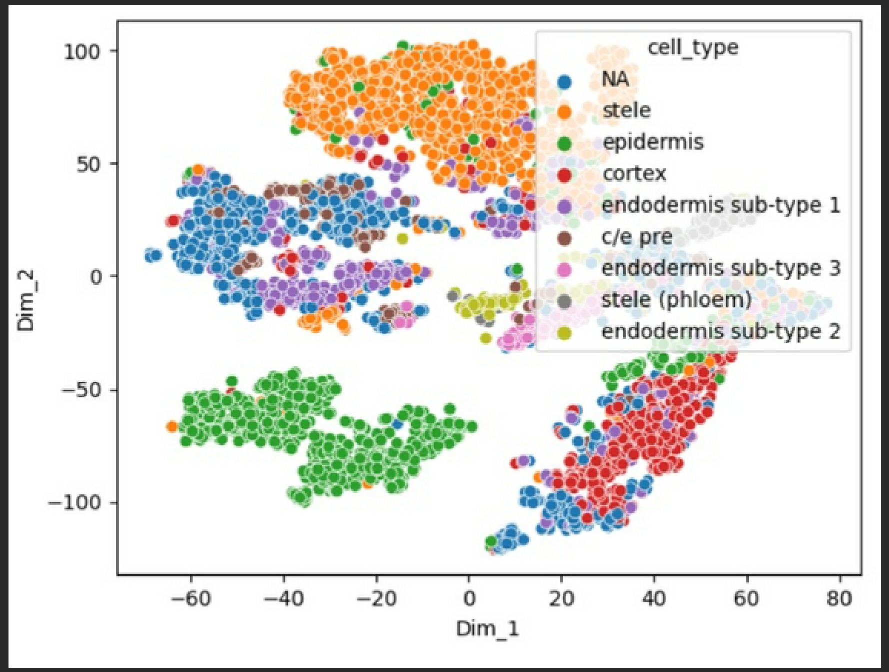
Hierarchical clustering



Dataset 2

1000 rows
12 features
Target: Profile

Prototype clustering



Dataset 3

5283rows
17 features
Target: Profile

DBSCAN

Q) 2b

Advantages

- Automatically discover arbitrarily shaped clusters when the algorithm is run.
- Find clusters completely surrounded by different clusters.
- It's Robust nature for the outlier detection makes it advantageous
- Require just 2 points, which are very insensitive to ordering the points in the database.
- It can automatically identify the noise data while clustering

Disadvantages

- This is not partitionable at multiprocessor systems
- It becomes tricky with datasets of alternating densities.
- Sensitive towards minPoints and EVS
- Fails to identify clusters if densities vary and if the data is too sparse.
- Sampling affects density measures.
- When the dataset is of neck type it fails.

Q) 2c

Density based clustering algorithm chosen is: **OPTICS**

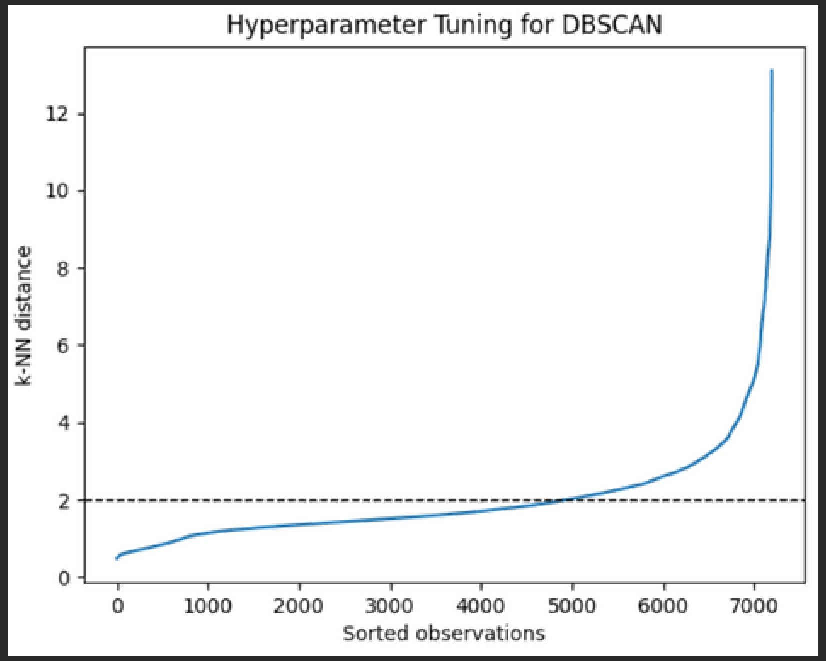
Advantages:

- It does not require density parameters.
- The clustering order is useful for extracting the basic clustering information.
- It operates on variable epsilon.

Q) 2d

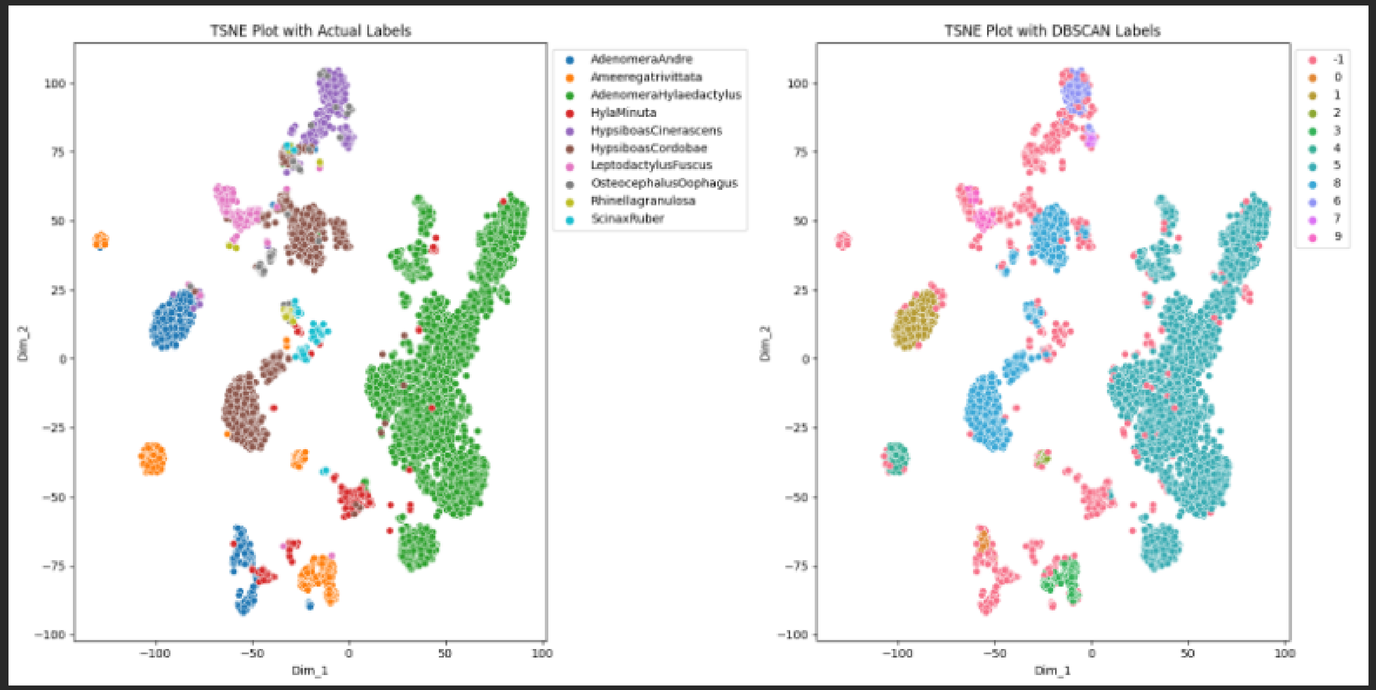
DBSCAN Implementation

Dataset 1



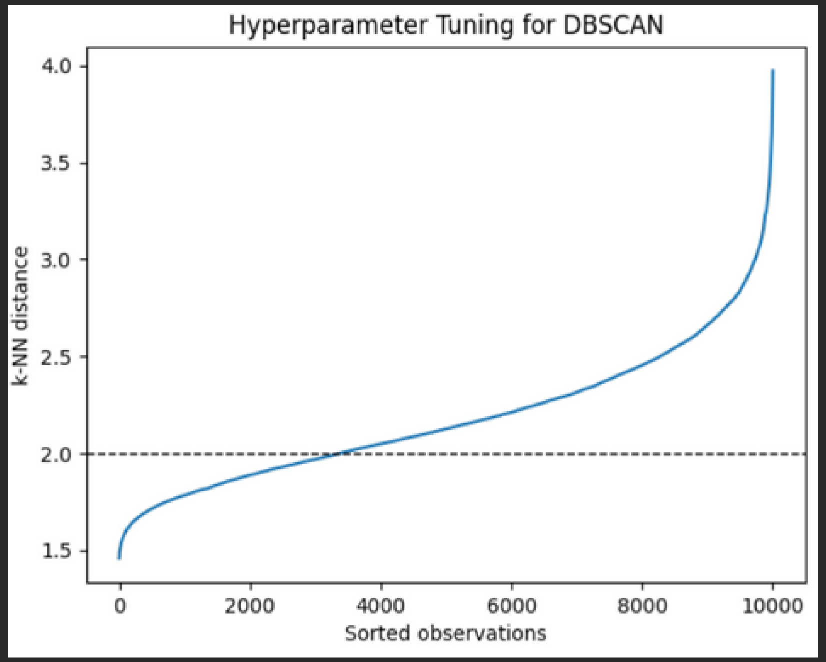
epsilon: 1.9

Silhouette Coefficient: 0.234
Davies Bouldin Score: 2.185
Adjusted Rand Index: 0.802
Adjusted Mutual Information: 0.698



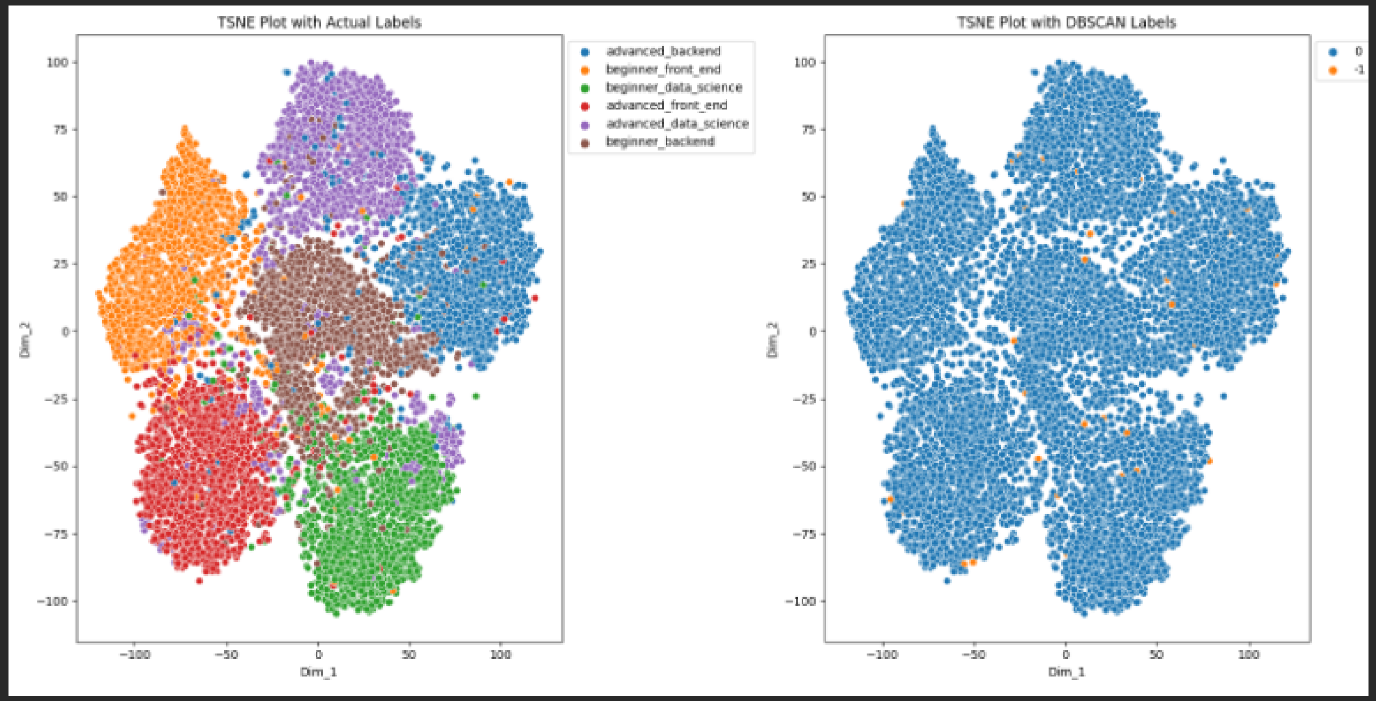
Est. Clusters: 10 Est. Noice Points: 1467

Dataset 2



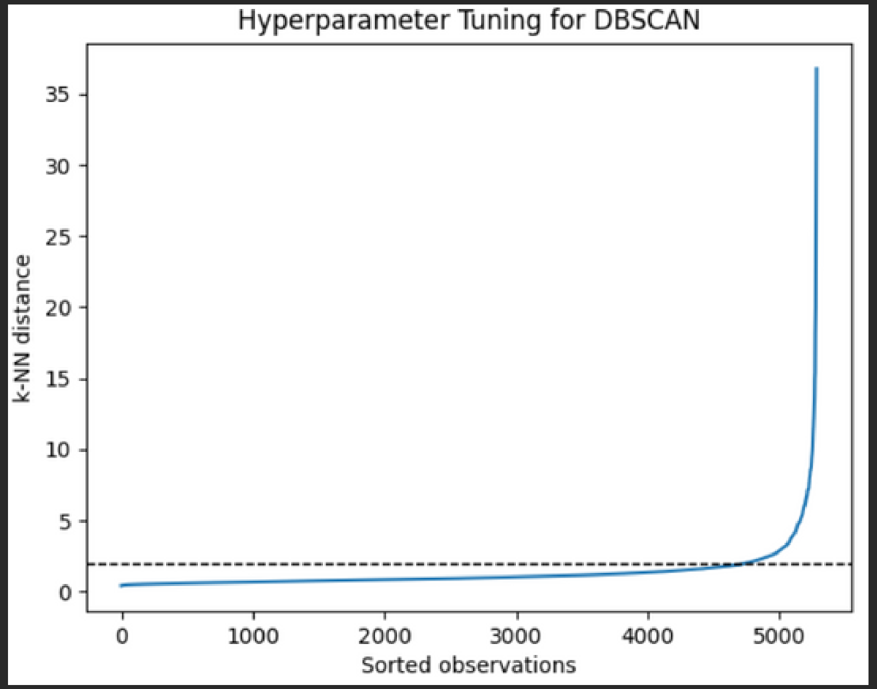
epsilon: 2.6

Silhouette Coefficient: 0.263
Davies Bouldin Score: 7.402
Adjusted Rand Index: 0.000
Adjusted Mutual Information: 0.000



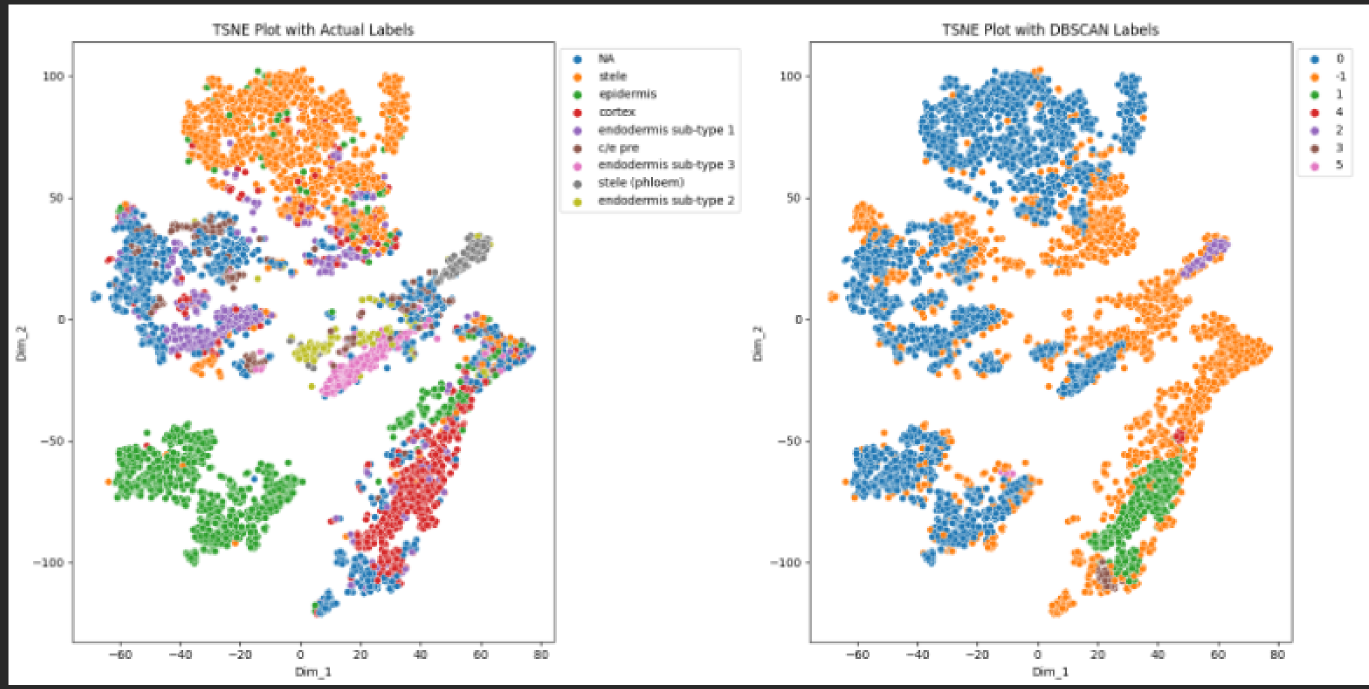
Est. Clusters: 1 Est. Noice Points: 51

Dataset 3



epsilon: 0.7

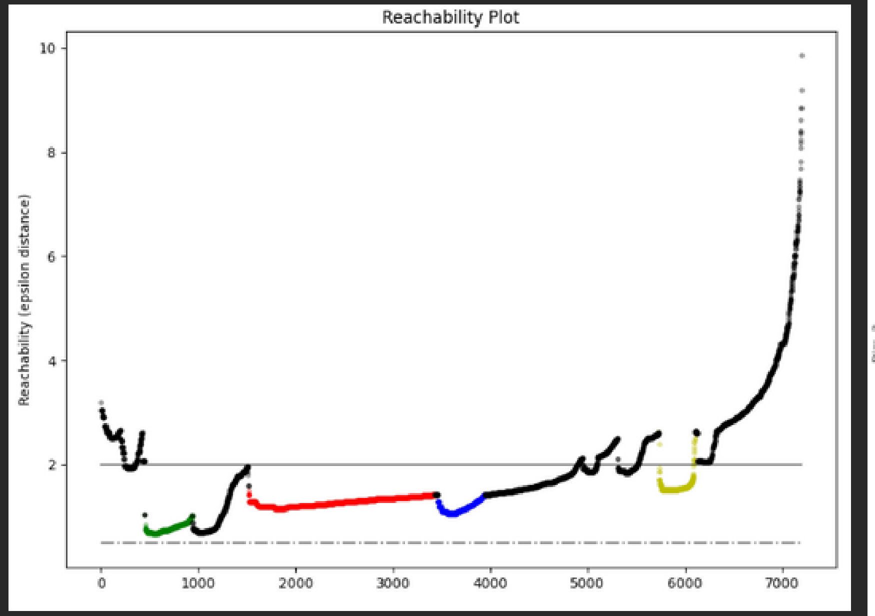
Silhouette Coefficient: -0.081
Davies Bouldin Score: 1.669
Adjusted Rand Index: 0.119
Adjusted Mutual Information: 0.206



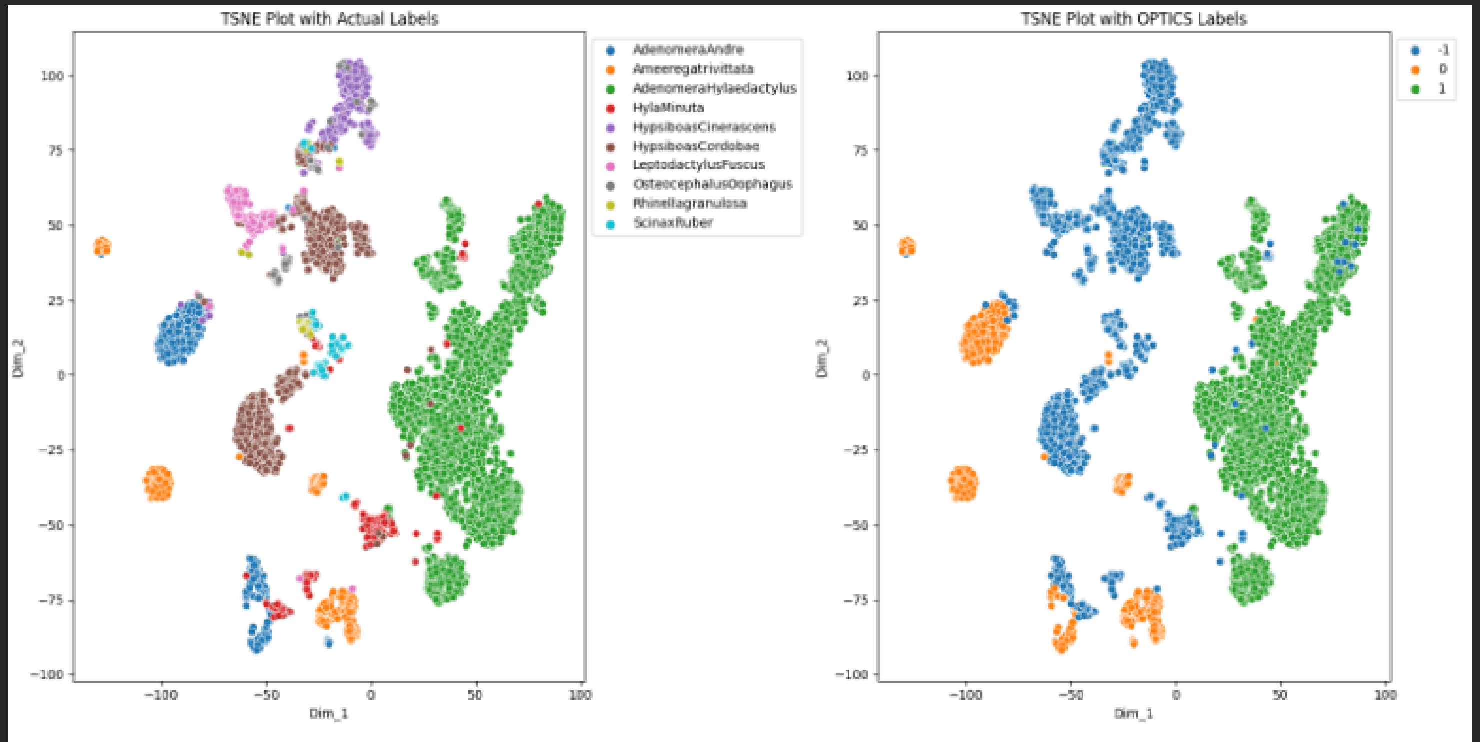
Est. Clusters: 6 Est. Noice Points: 1871

OPTICS Implementation

Dataset 1

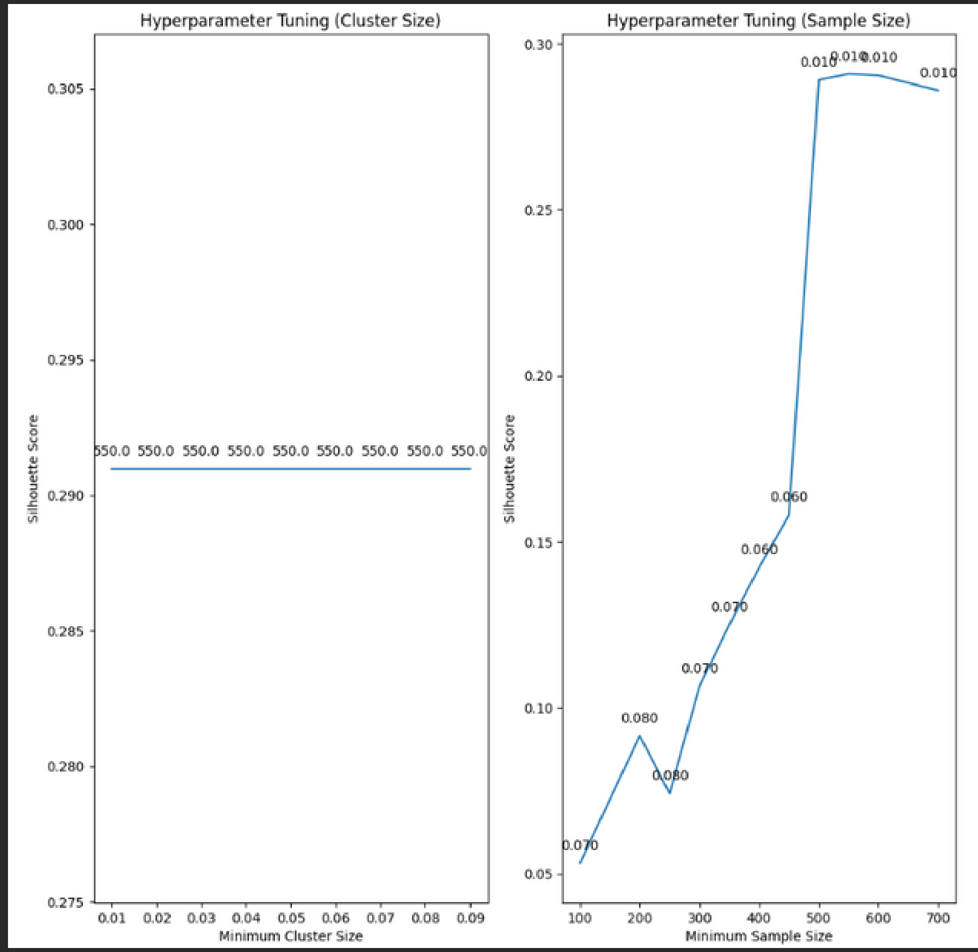


Reachability Plot

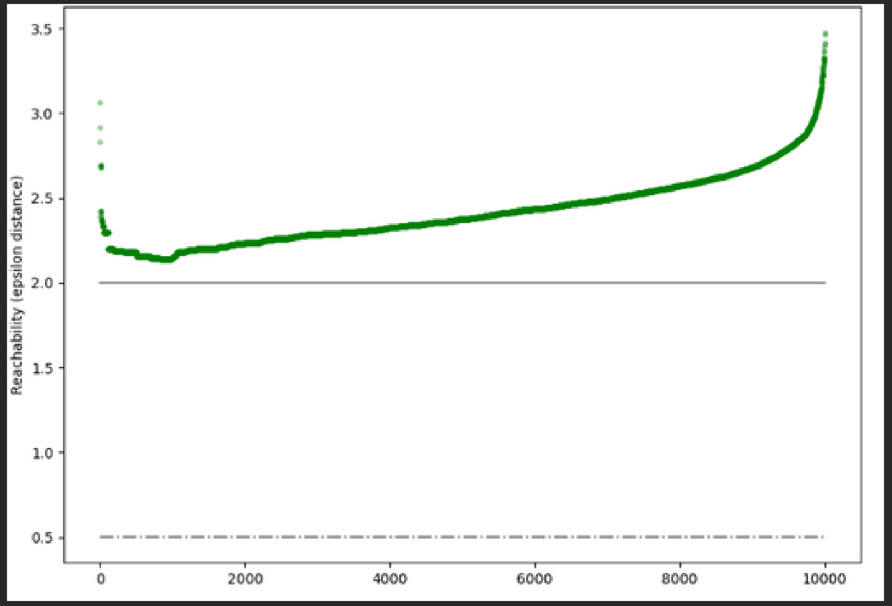


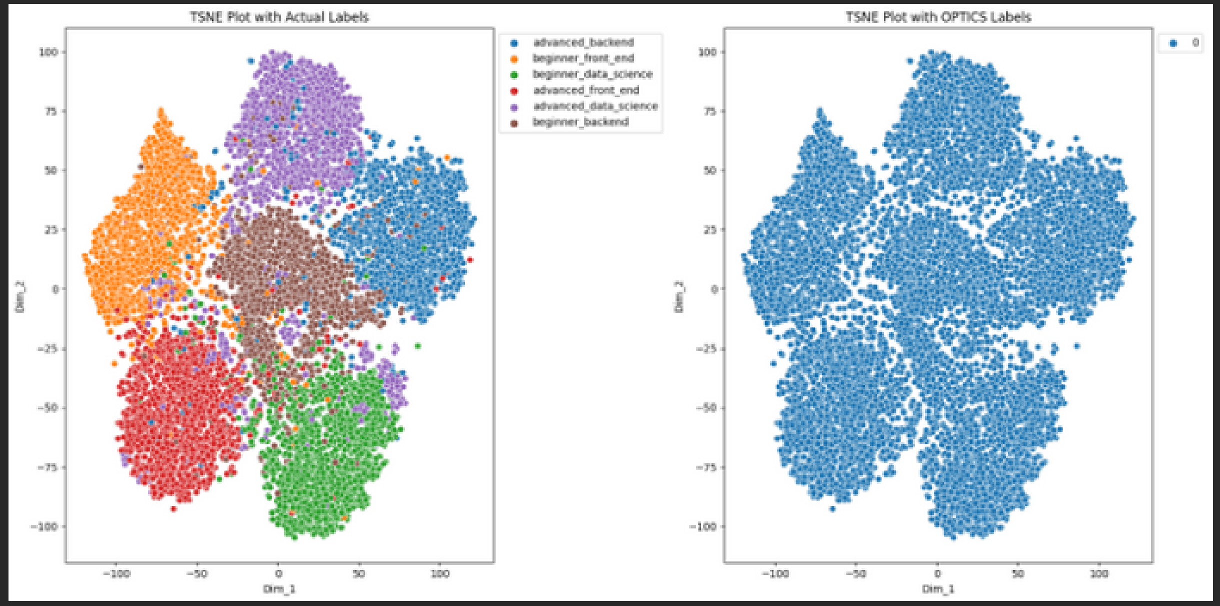
Optics Clustering Plot

Silhouette Coefficient: 0.2910
Davies Bouldin Score: 2.3854
Adjusted Rand Index: 0.7252
Adjusted Mutual Information: 0.6879



Dataset 2

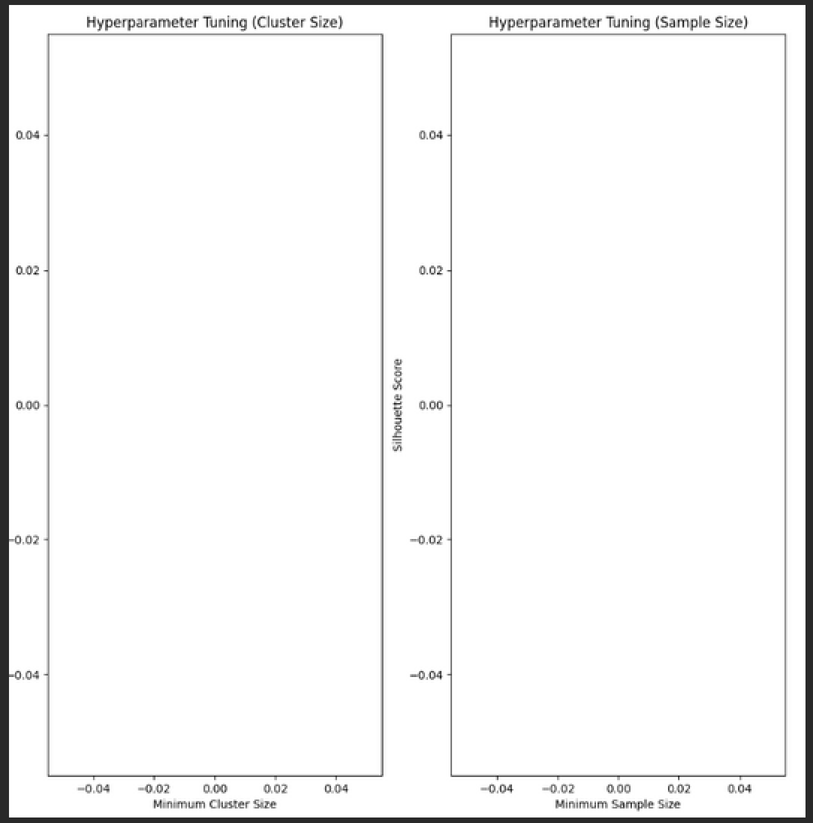




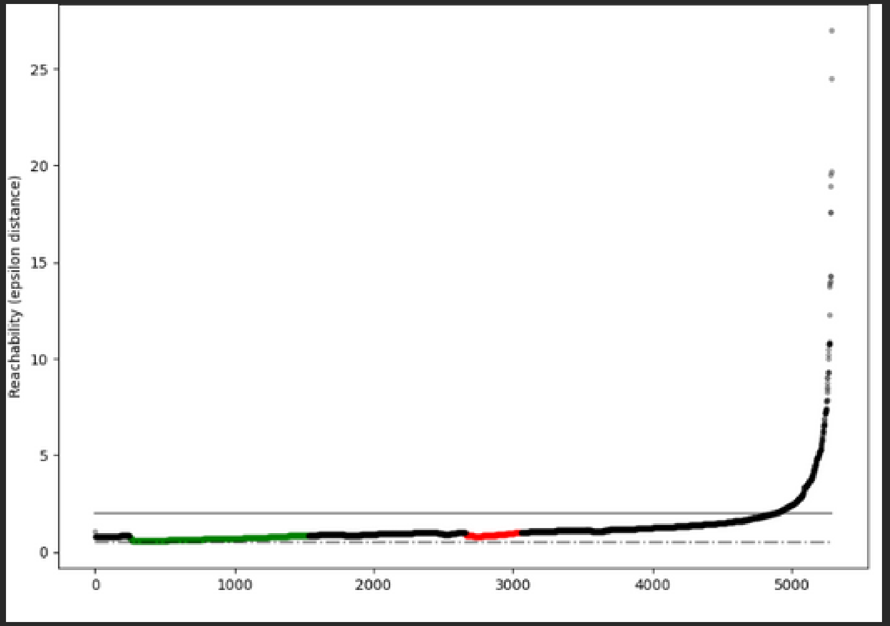
Reachability Plot

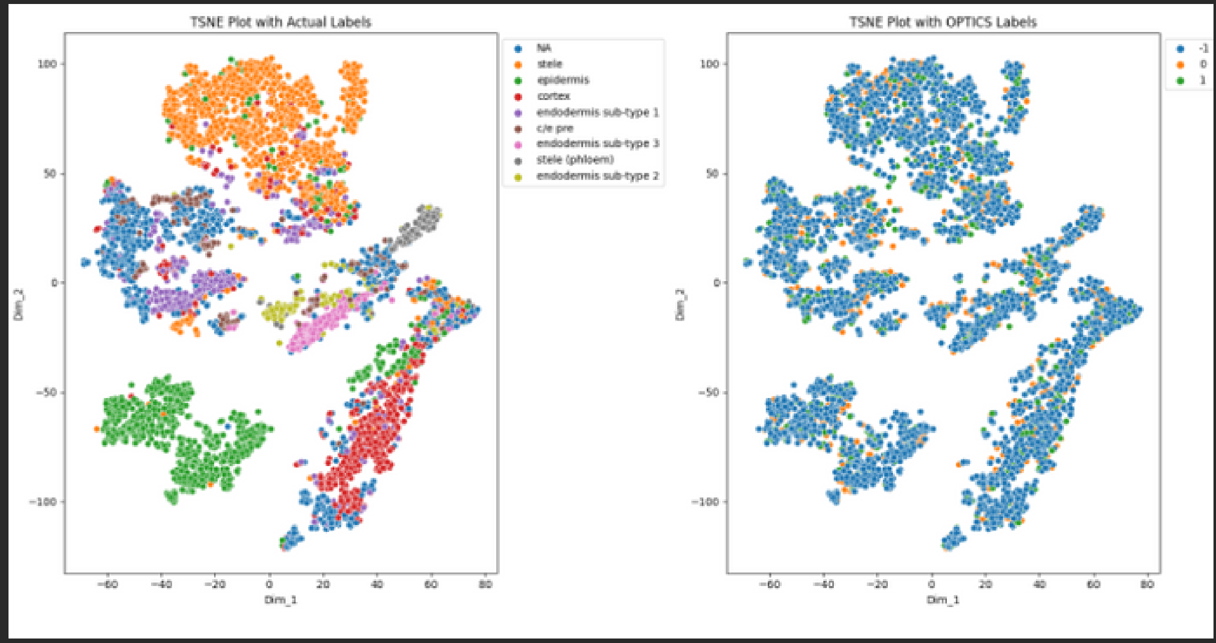
Optics Clustering Plot

Silhouette Coefficient: -inf
Davies Bouldin Score: inf
Adjusted Rand Index: 0.0000
Adjusted Mutual Information: 0.0000



Dataset 3

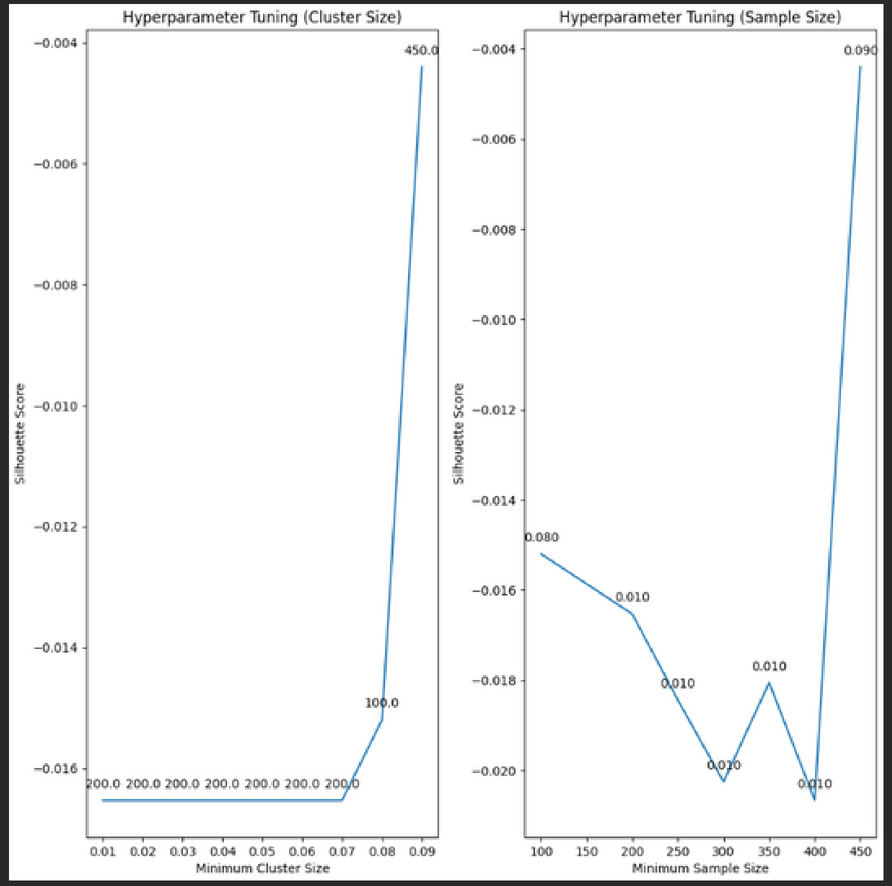




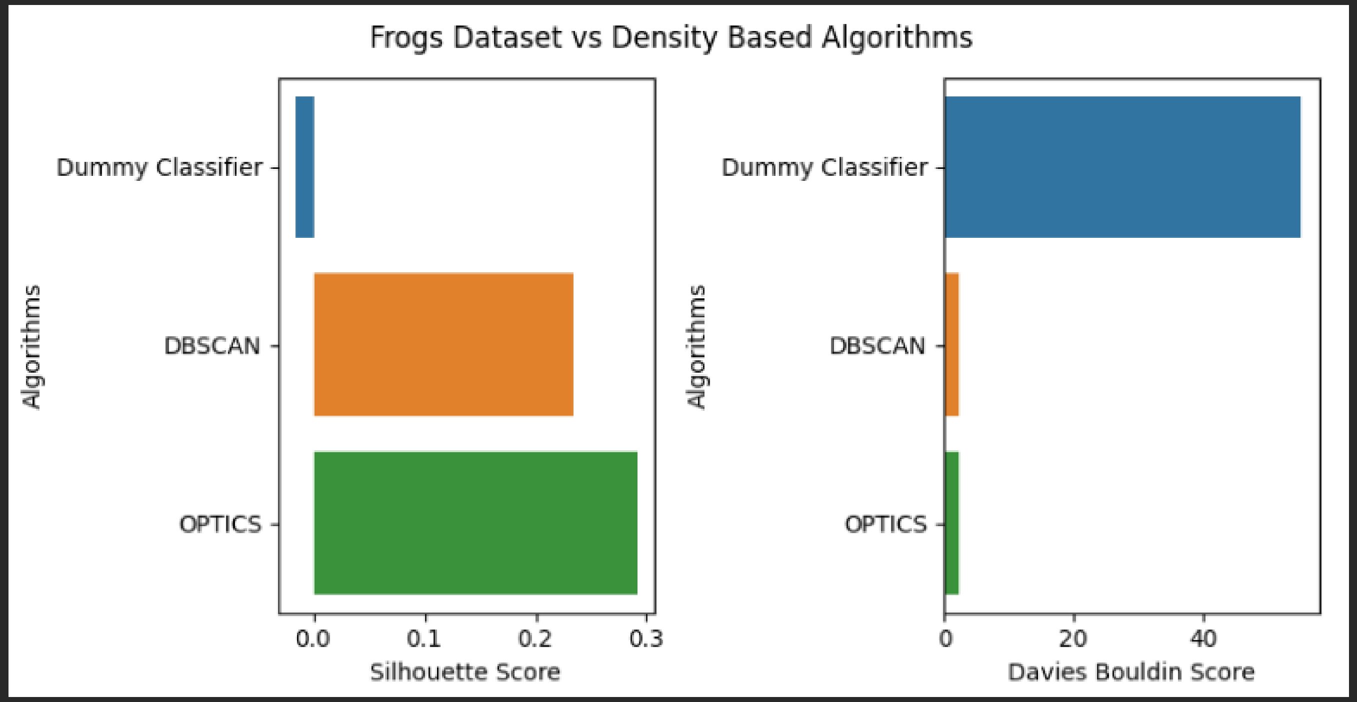
Reachability Plot

Optics Clustering Plot

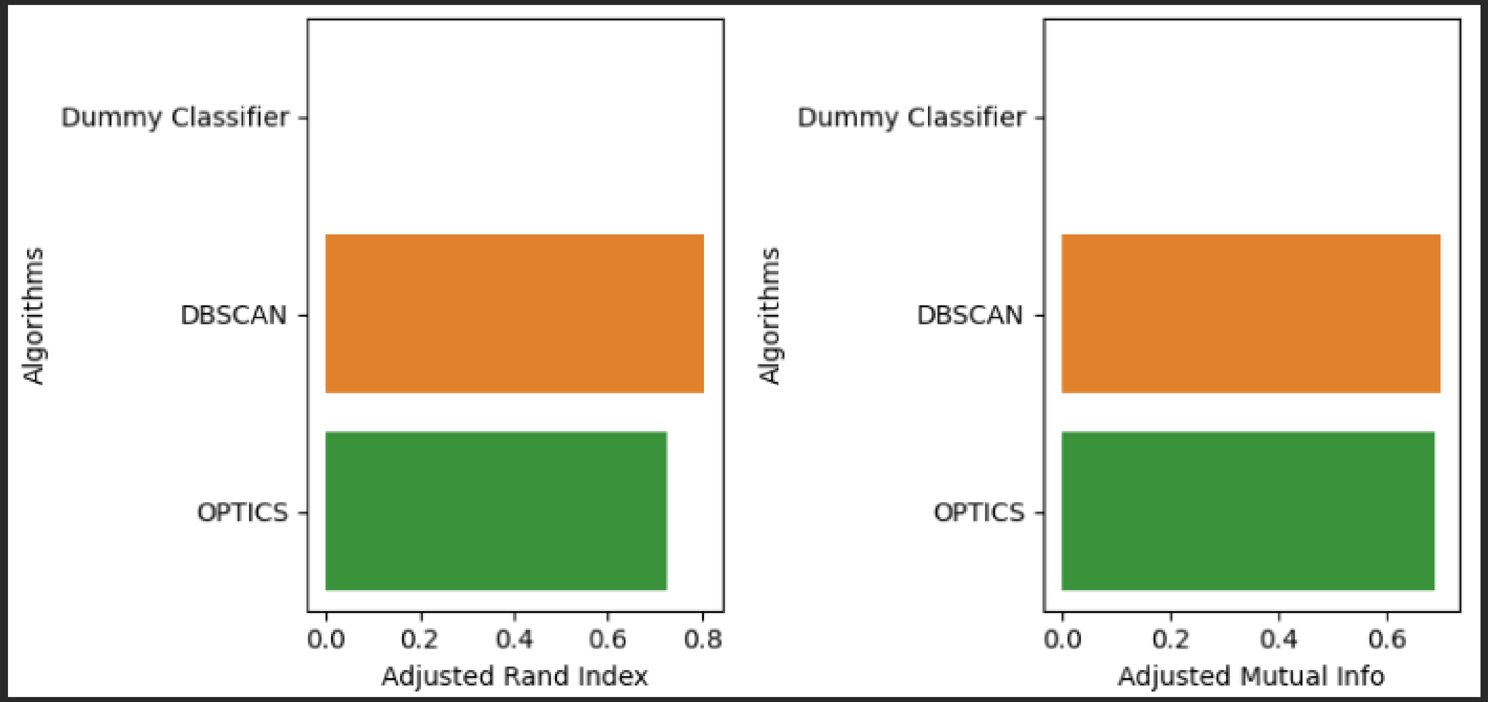
Silhouette Coefficient: -0.0152
Davies Bouldin Score: 39.68
Adjusted Rand Index: -0.0024
Adjusted Mutual Information: 0.0002



Q) 2e

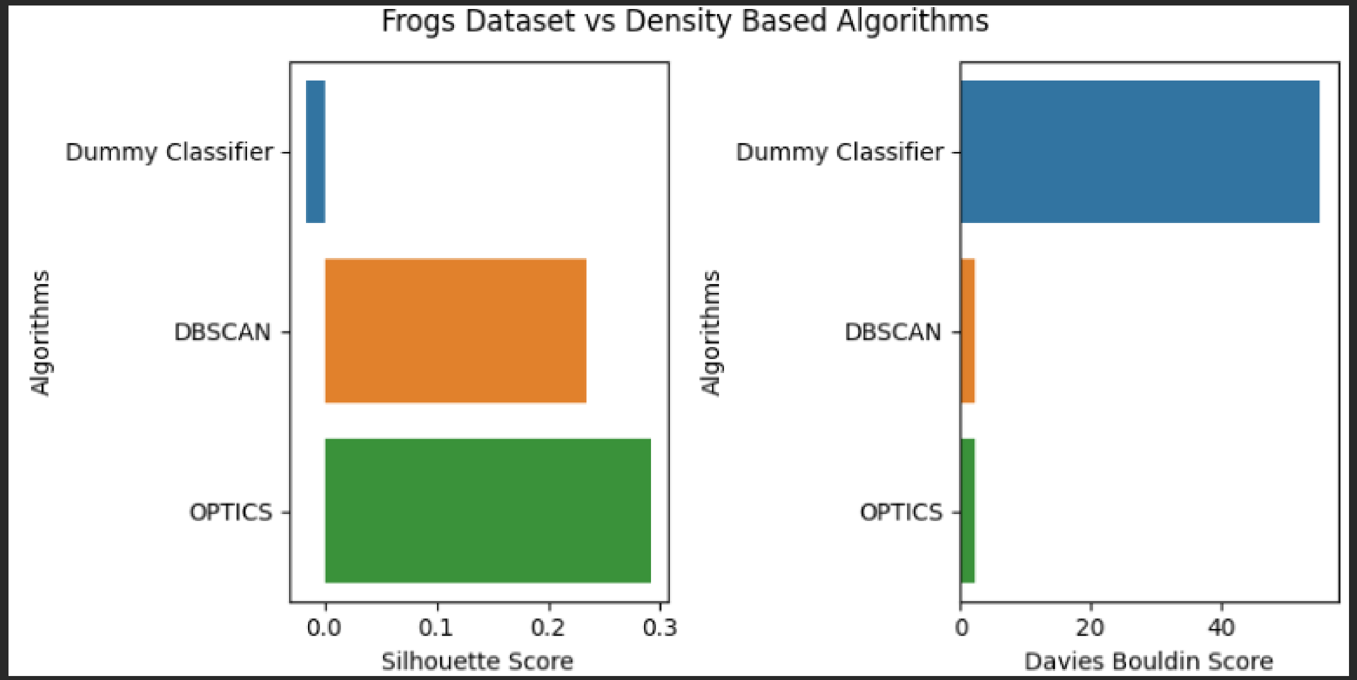


Frogs Dataset vs Density Based Algorithms
Dummy Classifier
DBSCAN
OPTICS
Algorithms
0.0
0.1
0.2
0.3
Silhouette Score
0
20
40
Davies Bouldin Score

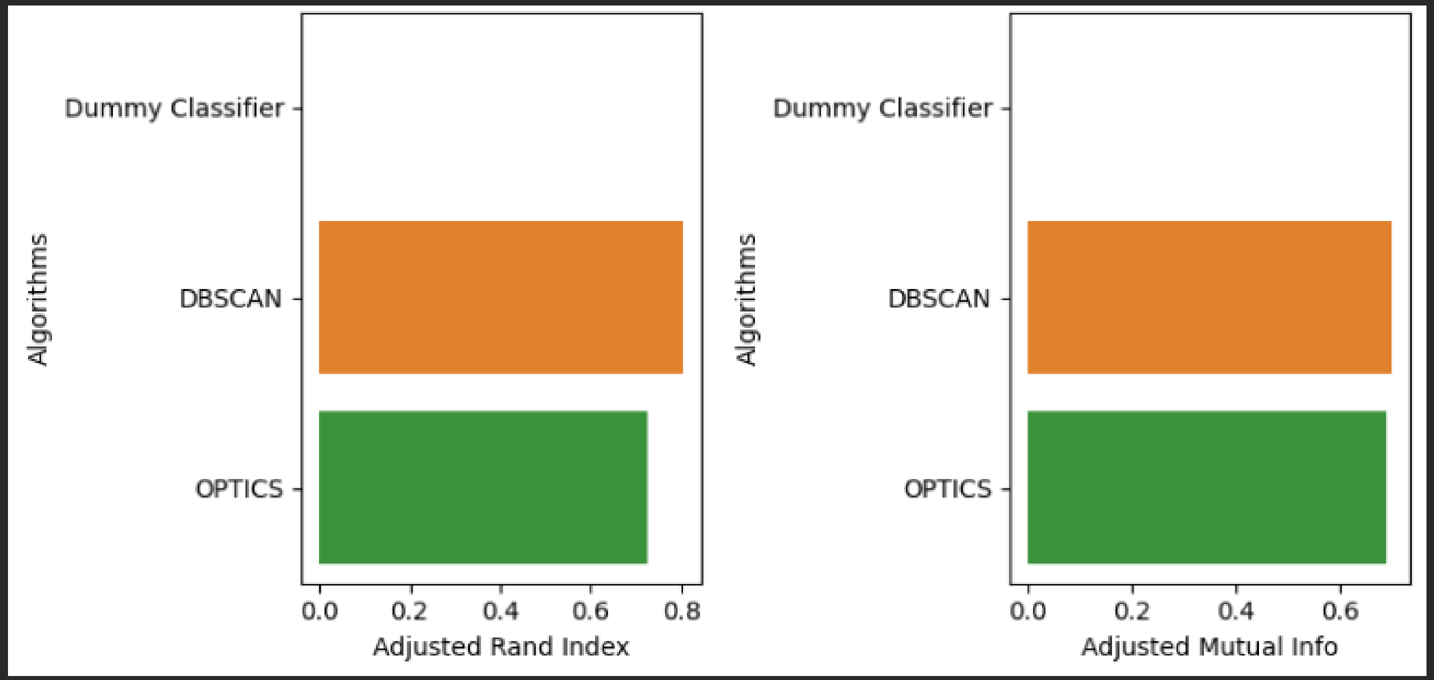


Dummy Classifier
DBSCAN
OPTICS
Algorithms
0.0
0.2
0.4
0.6
0.8
Adjusted Rand Index
Adjusted Mutual Info

Q) 2f

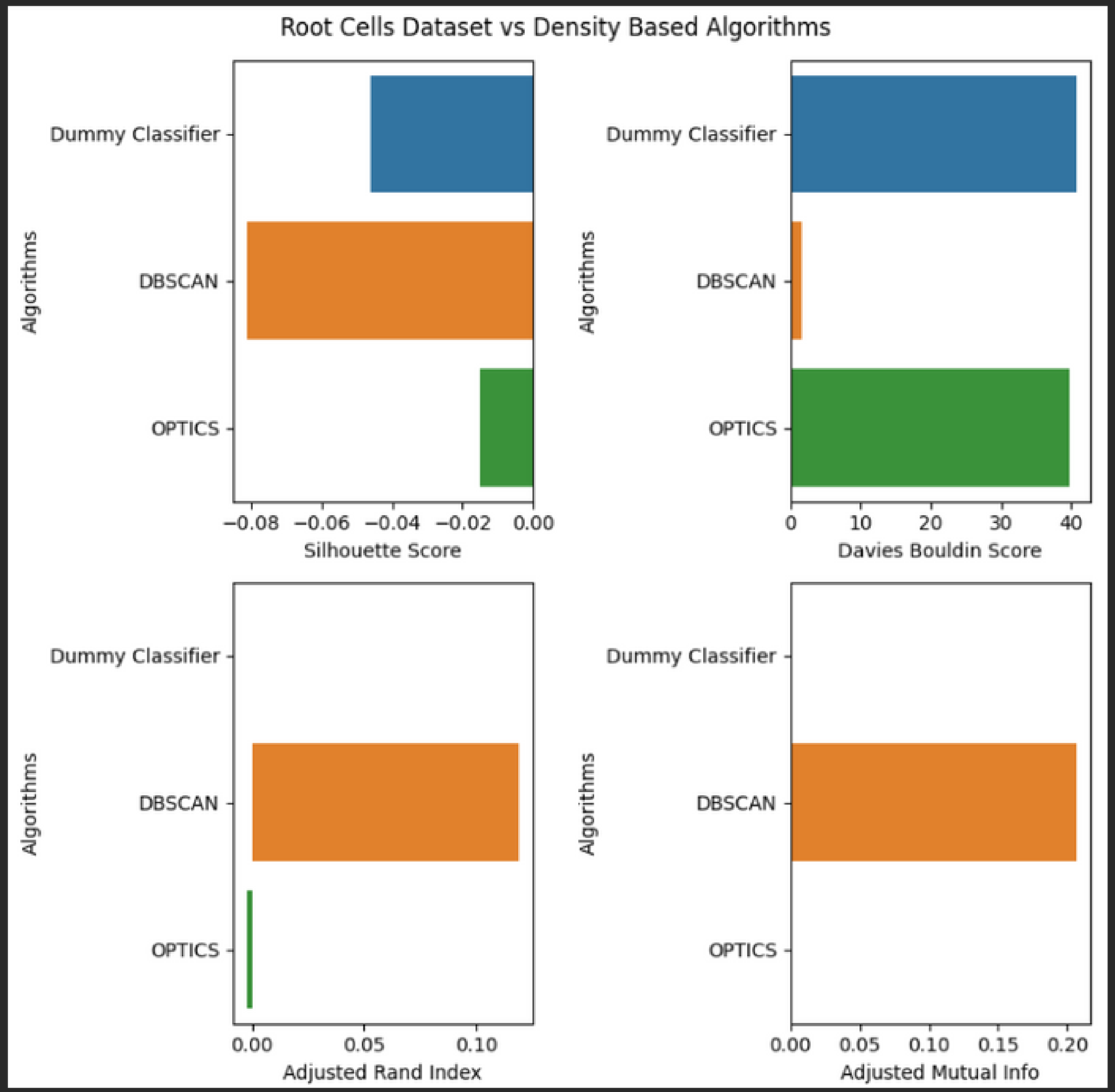


Frogs Dataset vs Density Based Algorithms
Dummy Classifier
DBSCAN
OPTICS
Algorithms
0.0
0.1
0.2
0.3
Silhouette Score
0
20
40
Davies Bouldin Score

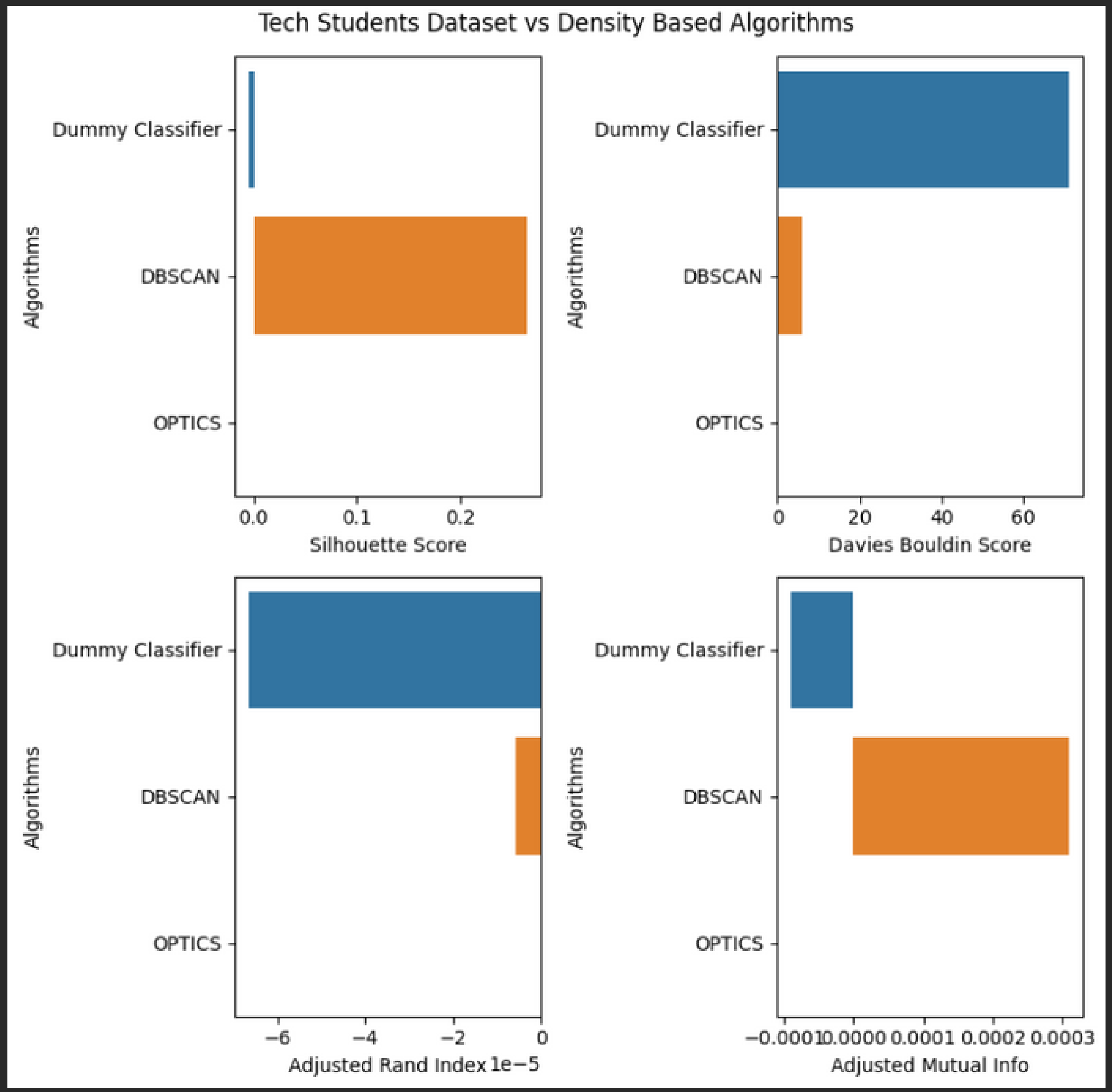


Dummy Classifier
DBSCAN
OPTICS
Algorithms
0.0
0.2
0.4
0.6
0.8
Adjusted Rand Index
Adjusted Mutual Info

Q) 2g

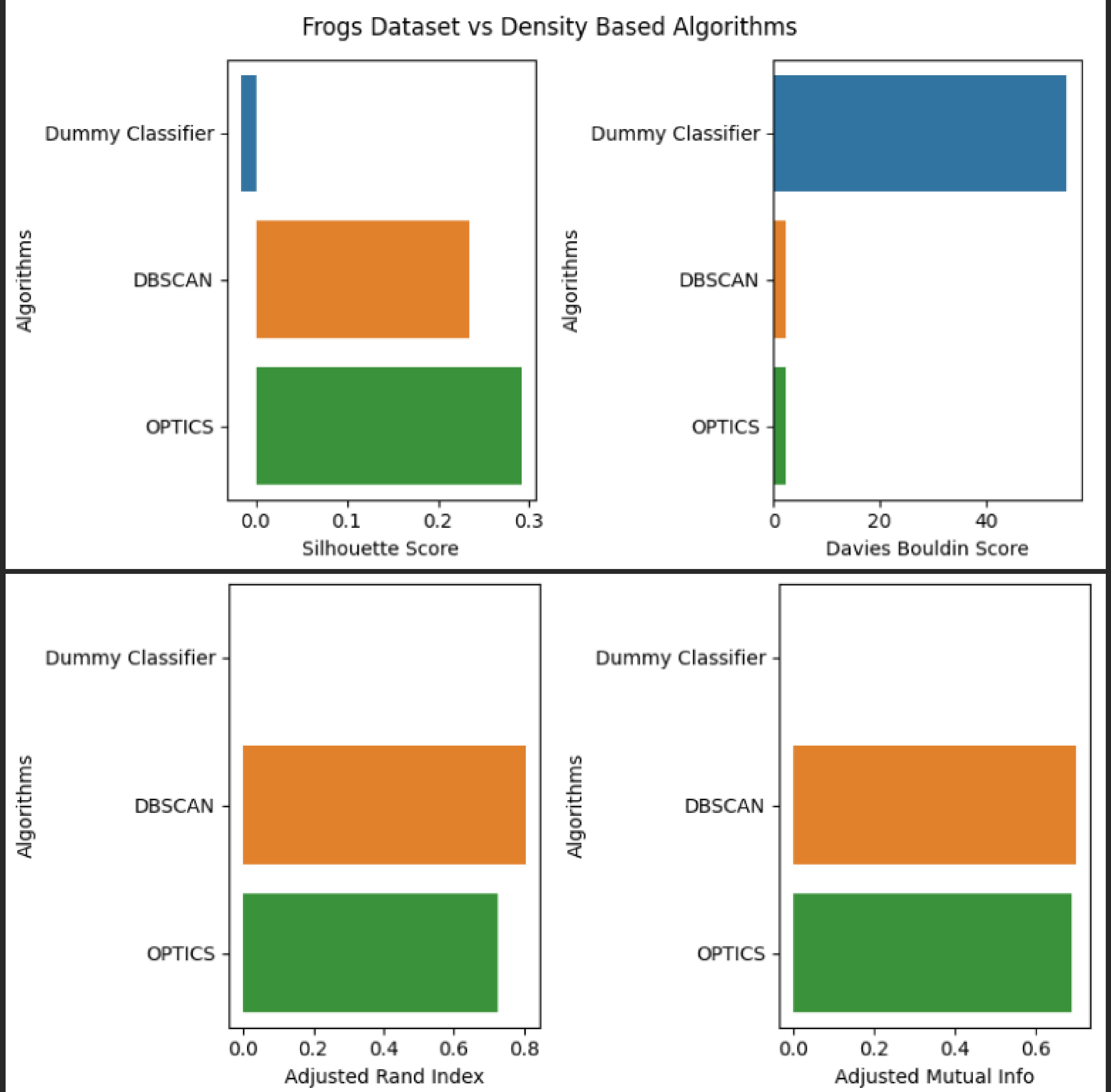


Root Cells Dataset vs Density Based Algorithms
Dummy Classifier
DBSCAN
OPTICS
Algorithms
Silhouette Score
Davies Bouldin Score
Adjusted Rand Index
Adjusted Mutual Info



Tech Students Dataset vs Density Based Algorithms
Dummy Classifier
DBSCAN
OPTICS
Algorithms
Silhouette Score
Davies Bouldin Score
Adjusted Rand Index
1e−5
Adjusted Mutual Info

RESULTS



Dataset 1 was the best fitting dataset in Q2.
According to the plotted graphs(As shown on the left side) DBSCAN showed slightly better results than Optics

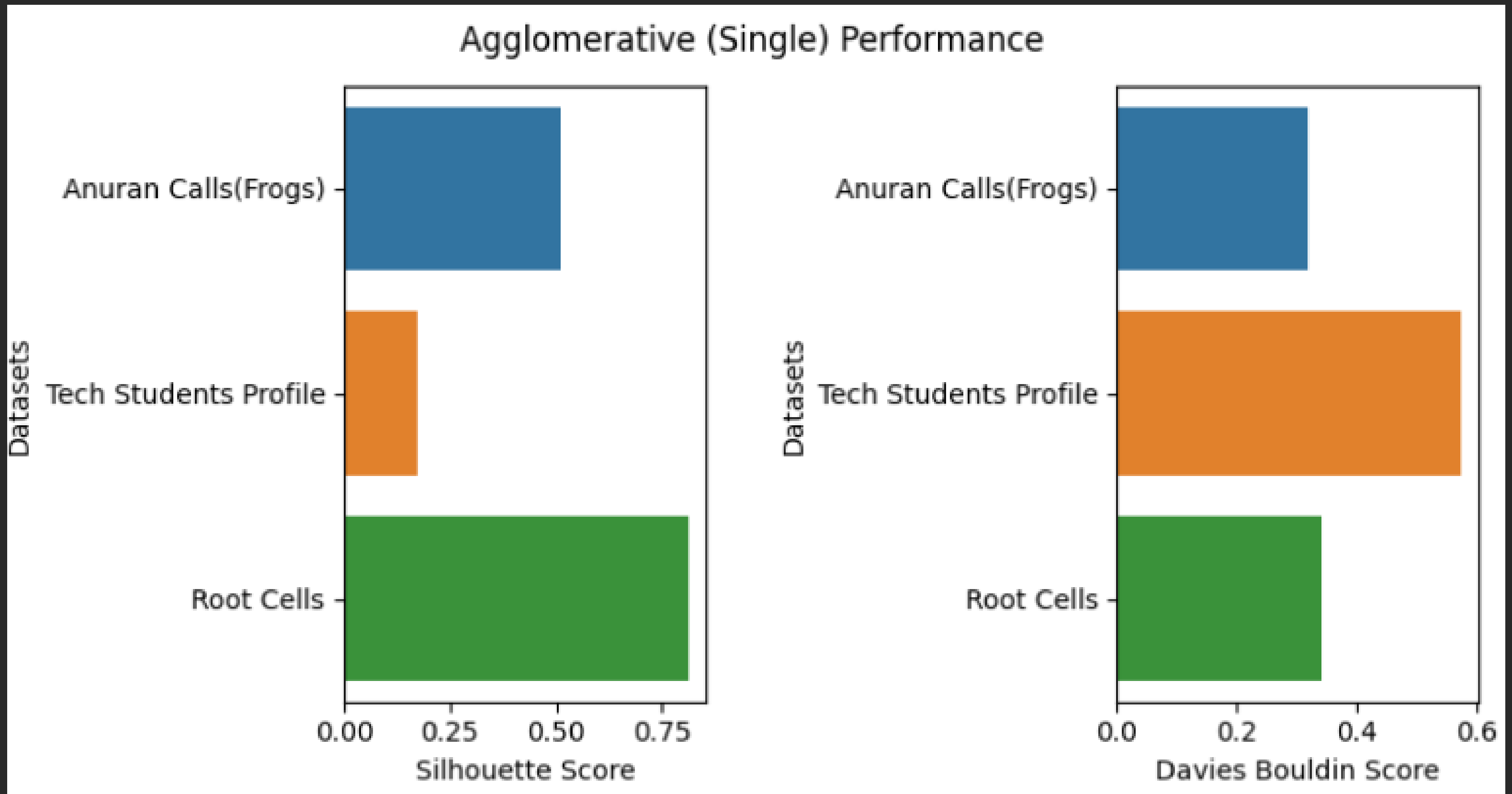
Q) 3 Hierarchical clustering

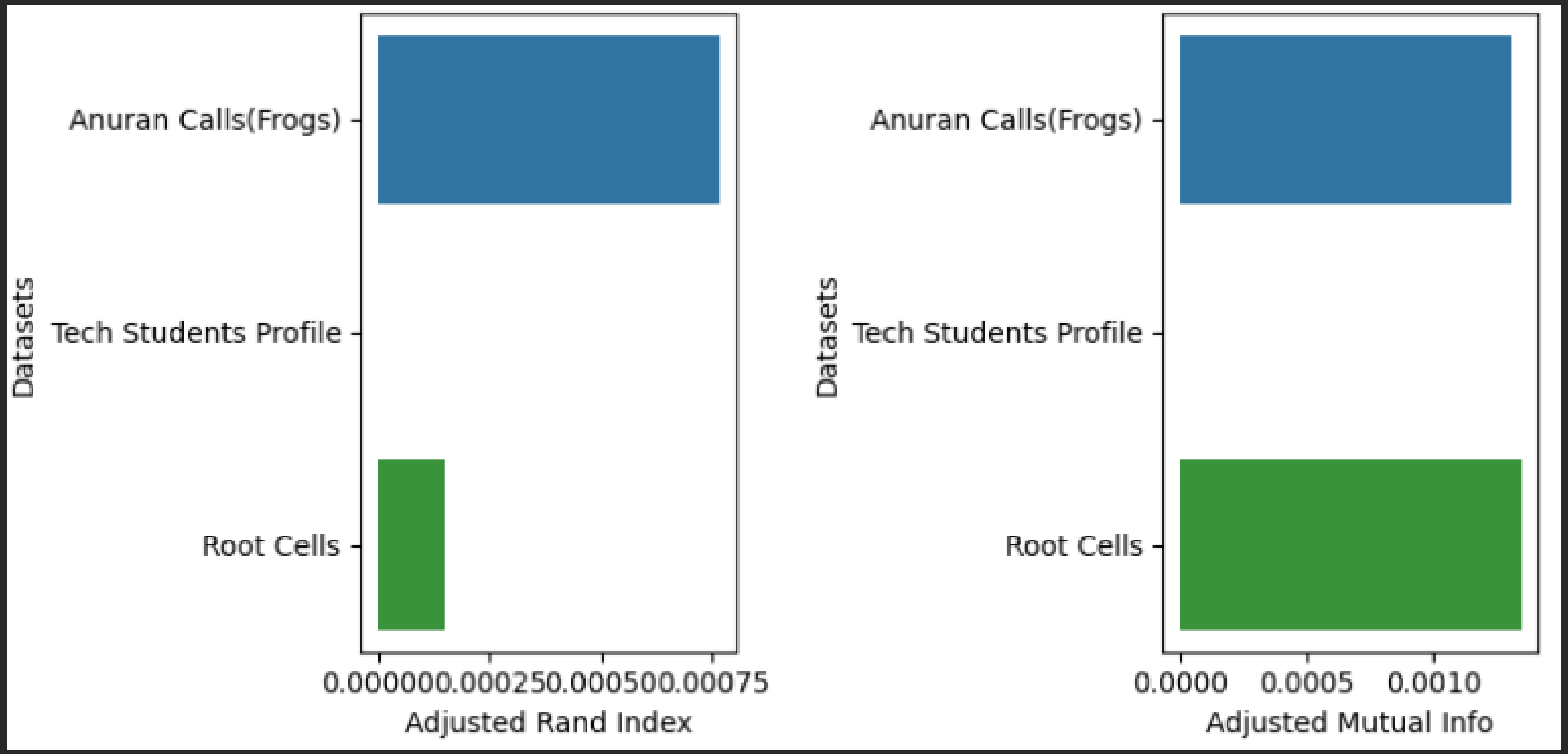
Two approaches : top-down(Agglomerative clustering) and bottom-up (Divisive Clustering)

Advantages

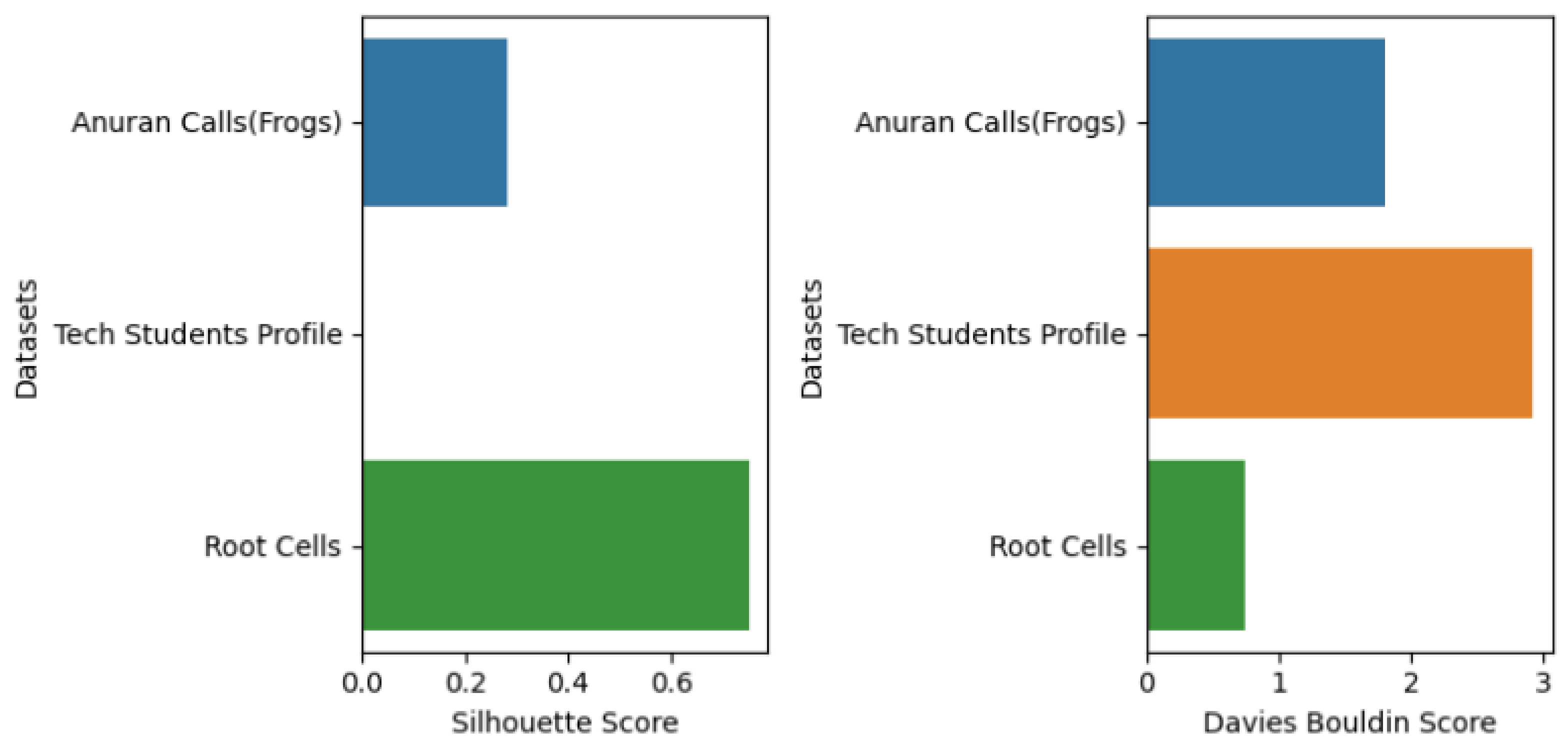
- There is no need to give any prior information about the number of clusters
- This algorithm is easy to implement as it gives best result.
- It works form the dissimilarities between the objects to be grouped together.

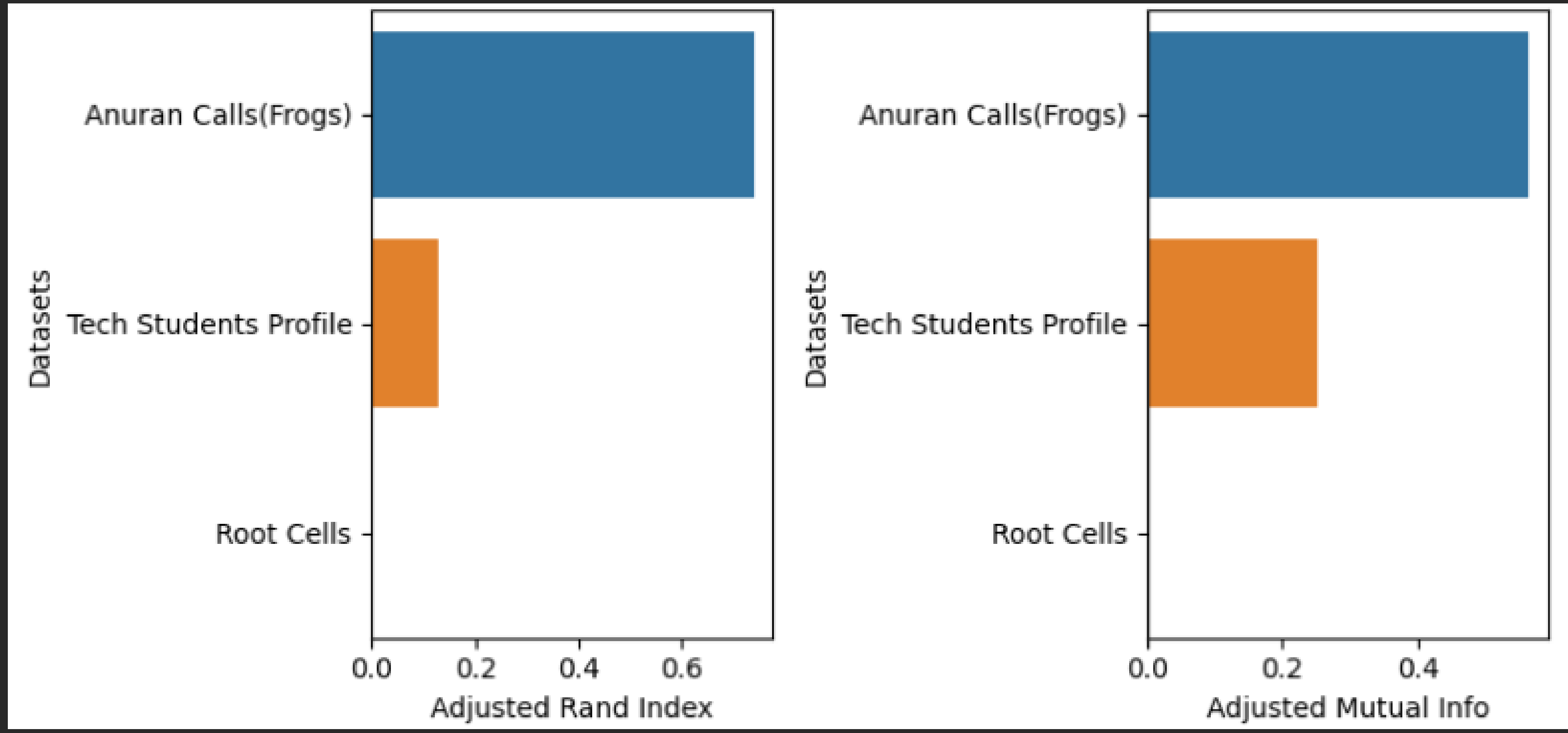
Single linkage



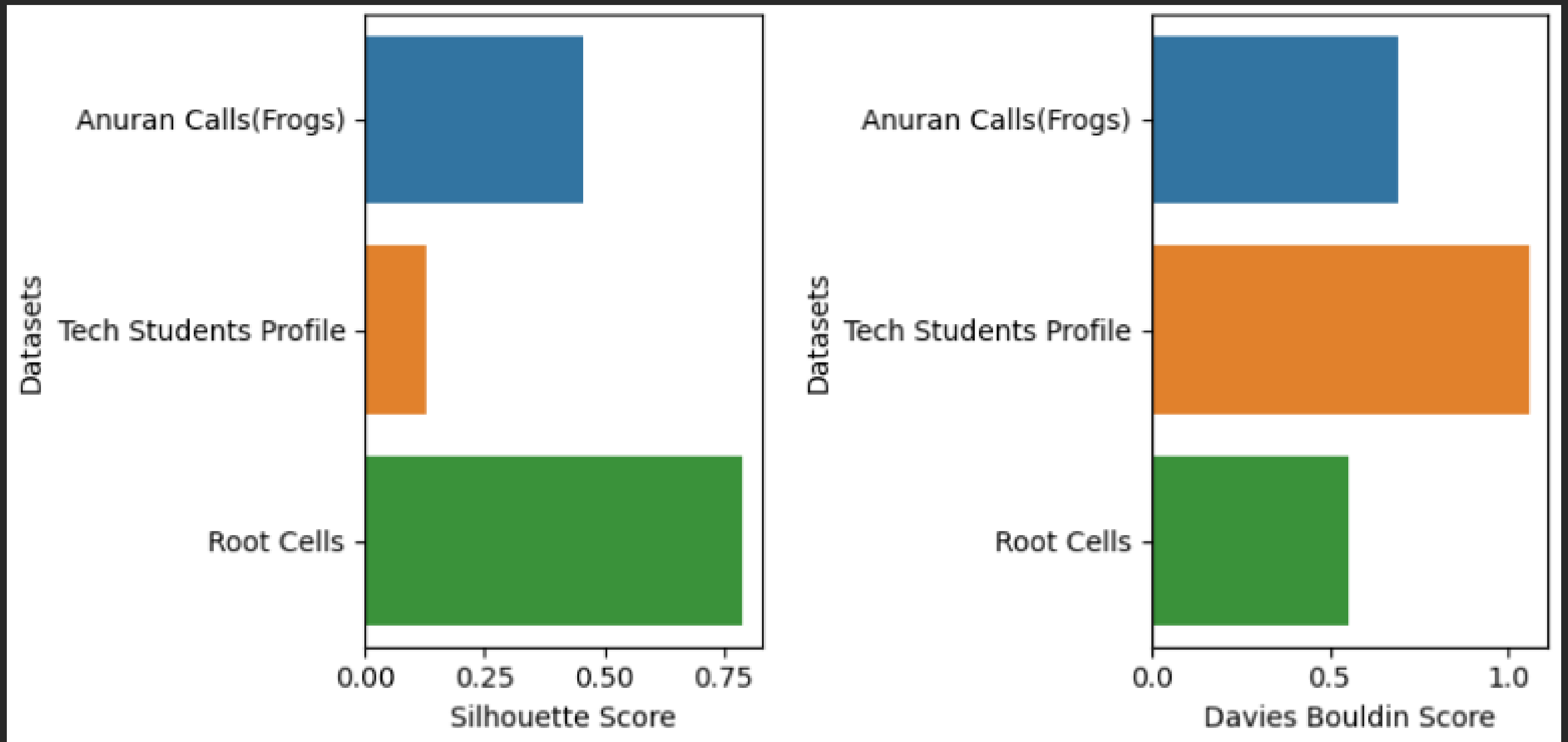


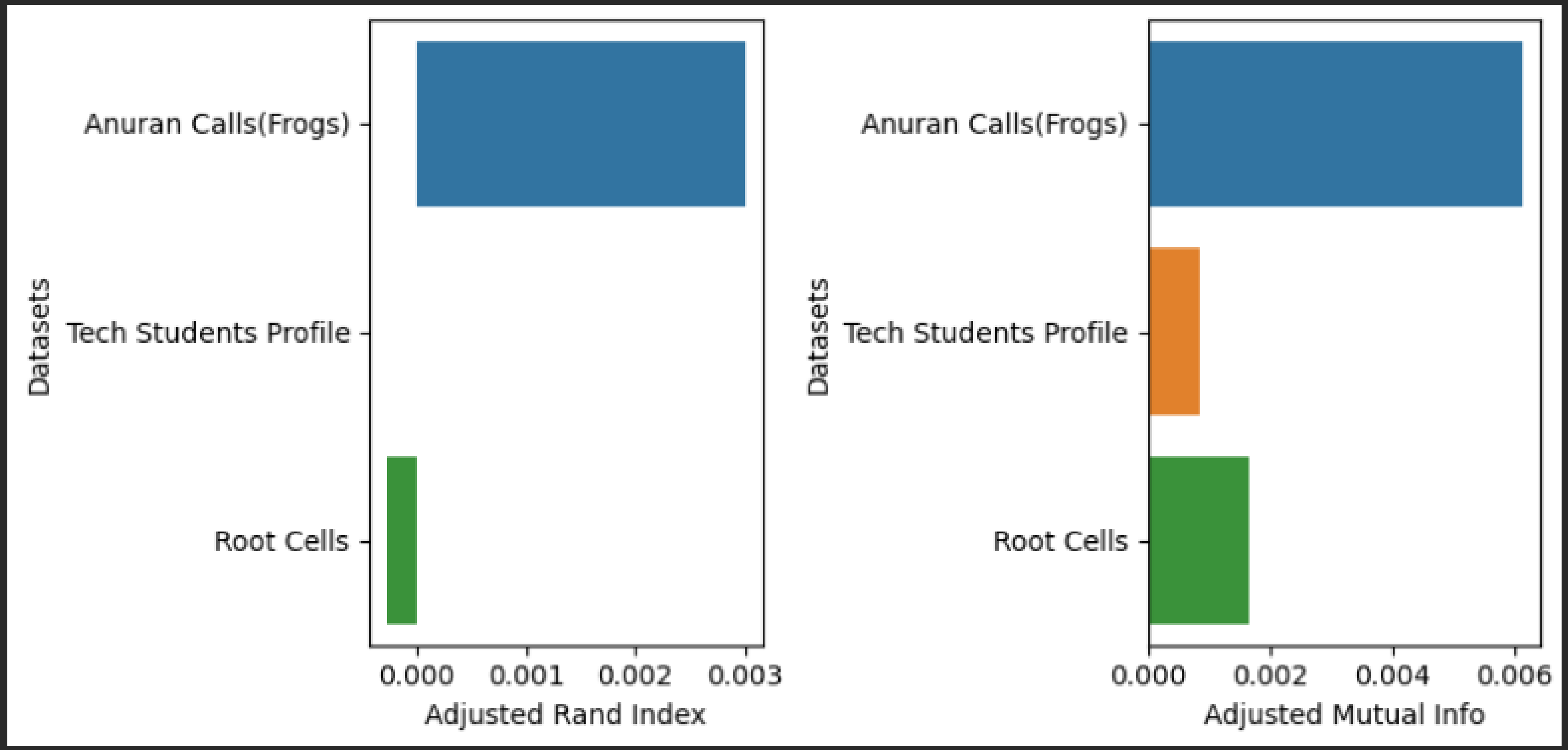
Complete linkage



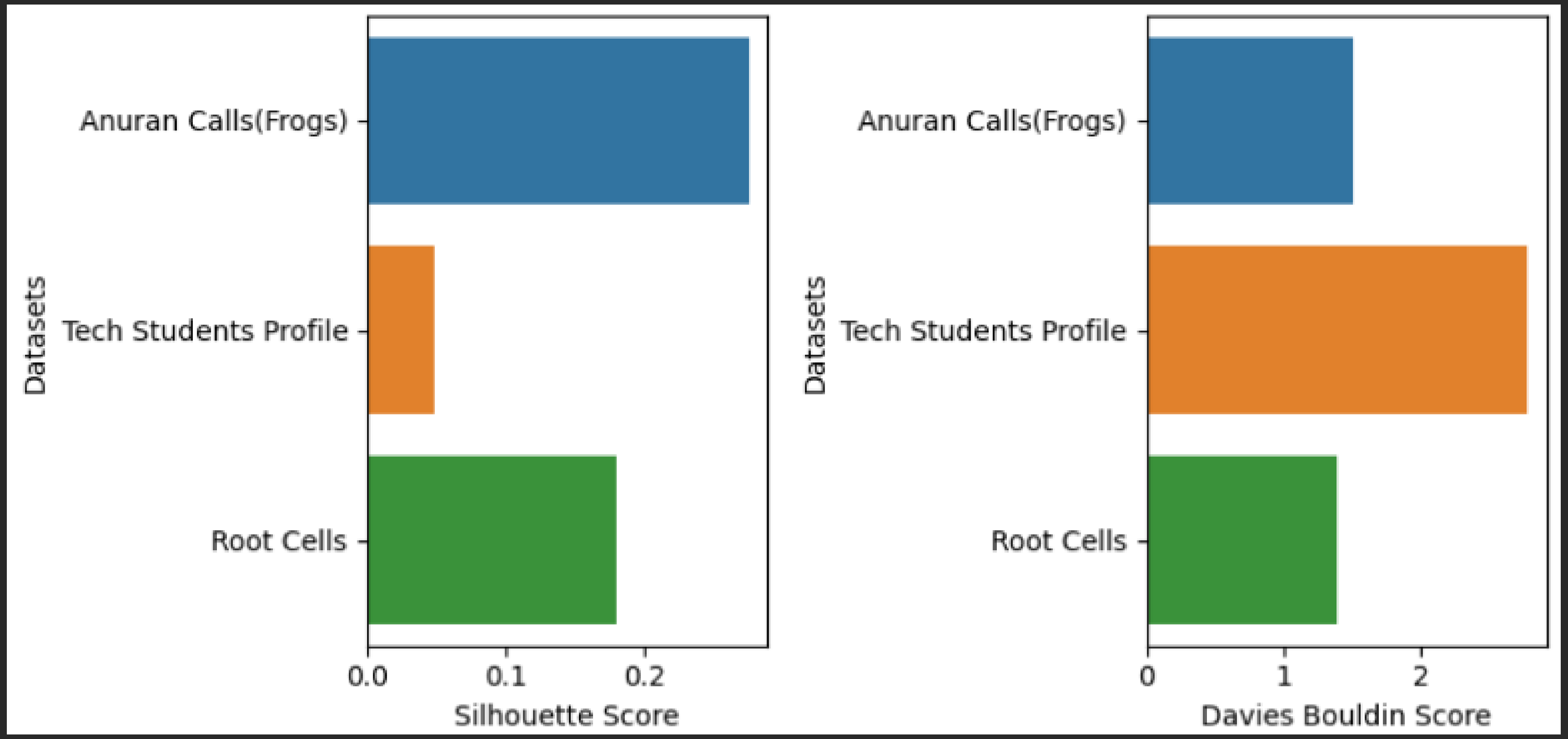


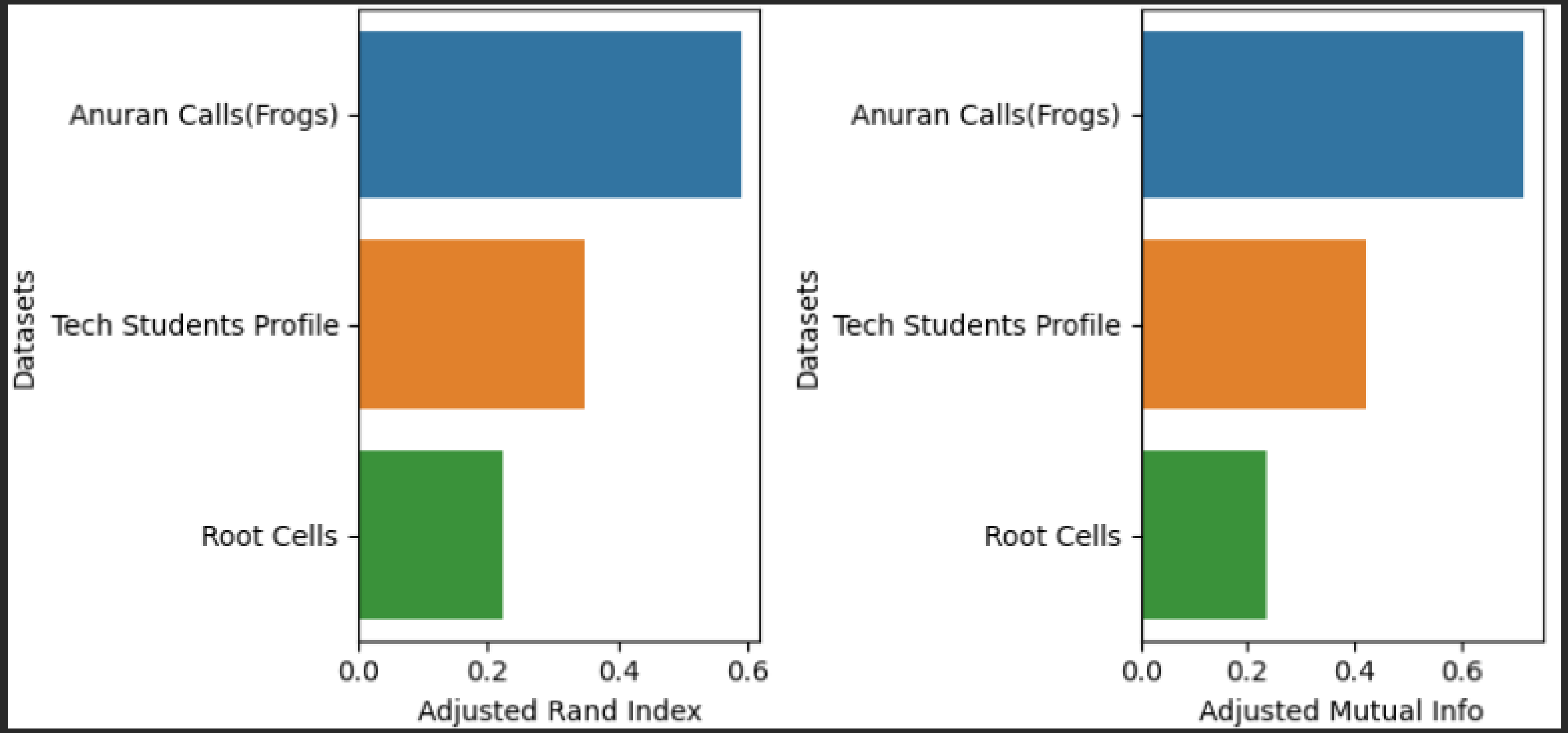
Average linkage





Ward linkage





RESULTS

The best shown results were from the complete linkages and the wards linkage. Though wards linkage showed the best results in this case.

The best results were shown by the frog dataset, though tech students dataset showed appreciable results in it too.

Q) 4 Prototype based clustering

Limitations of k-means:

- Clustering outliers
- Scaling with number of dimensions

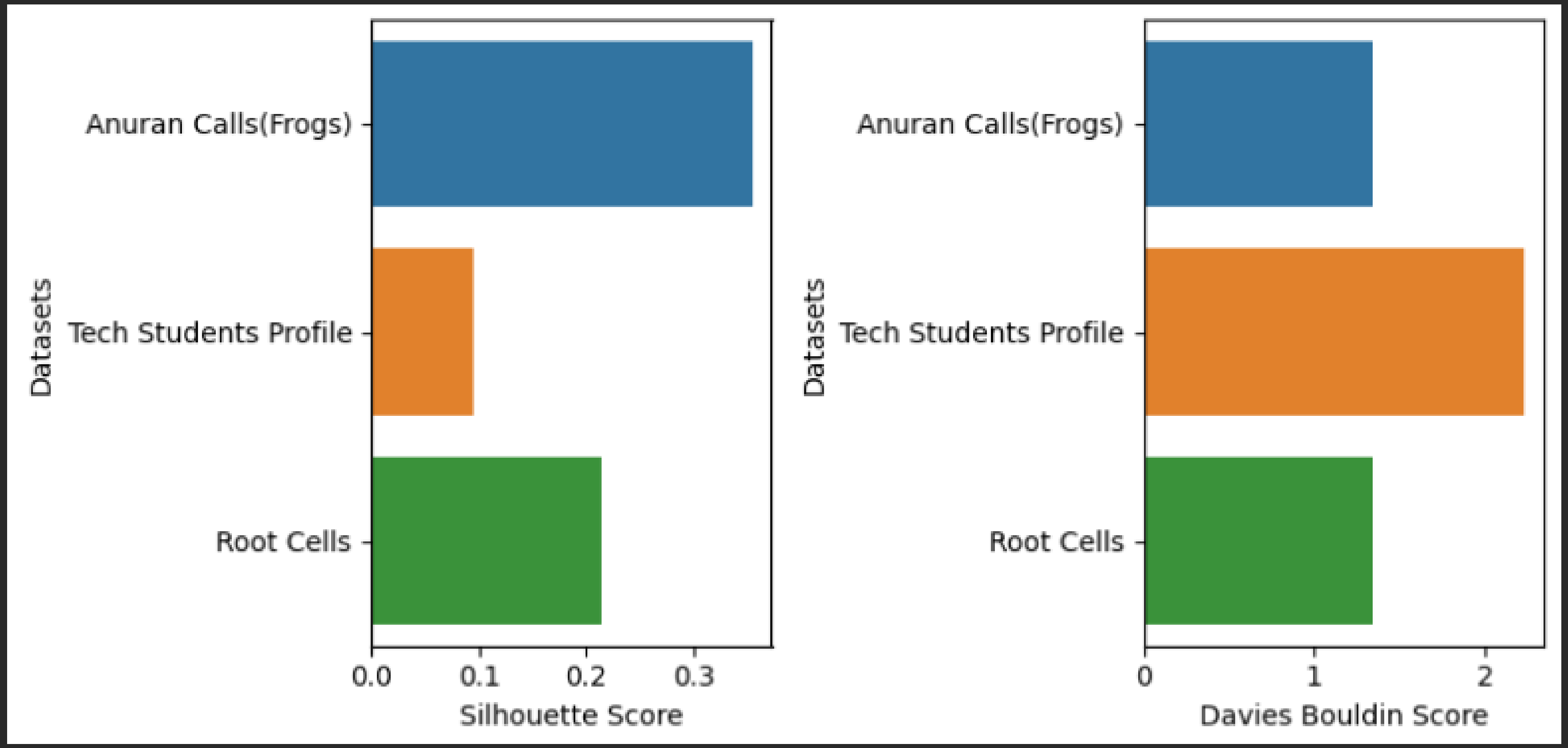
To encounter Clustering outliers limitation we use :

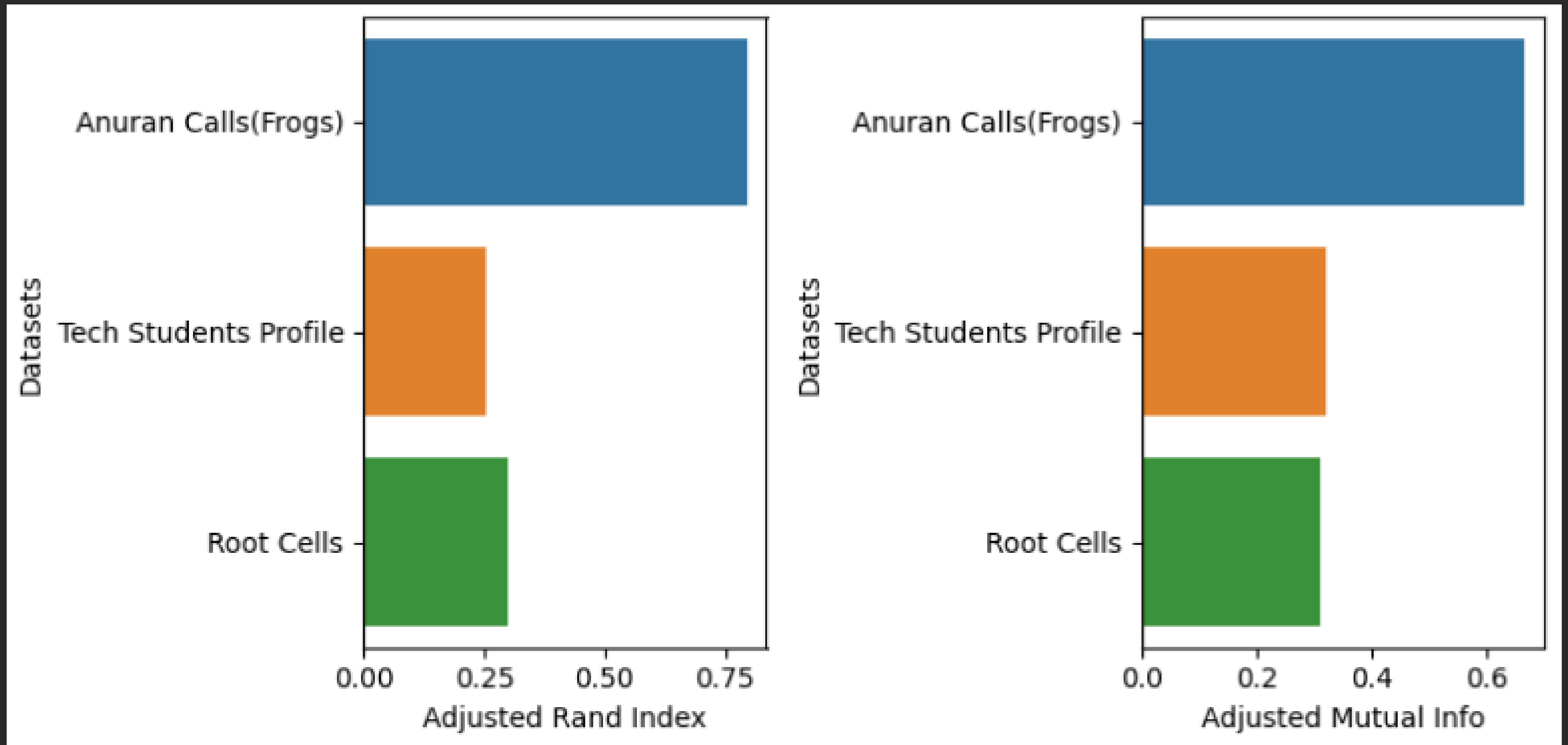
Clustering Outliers: K- medoids clustering

To encounter scaling limitation we can use:

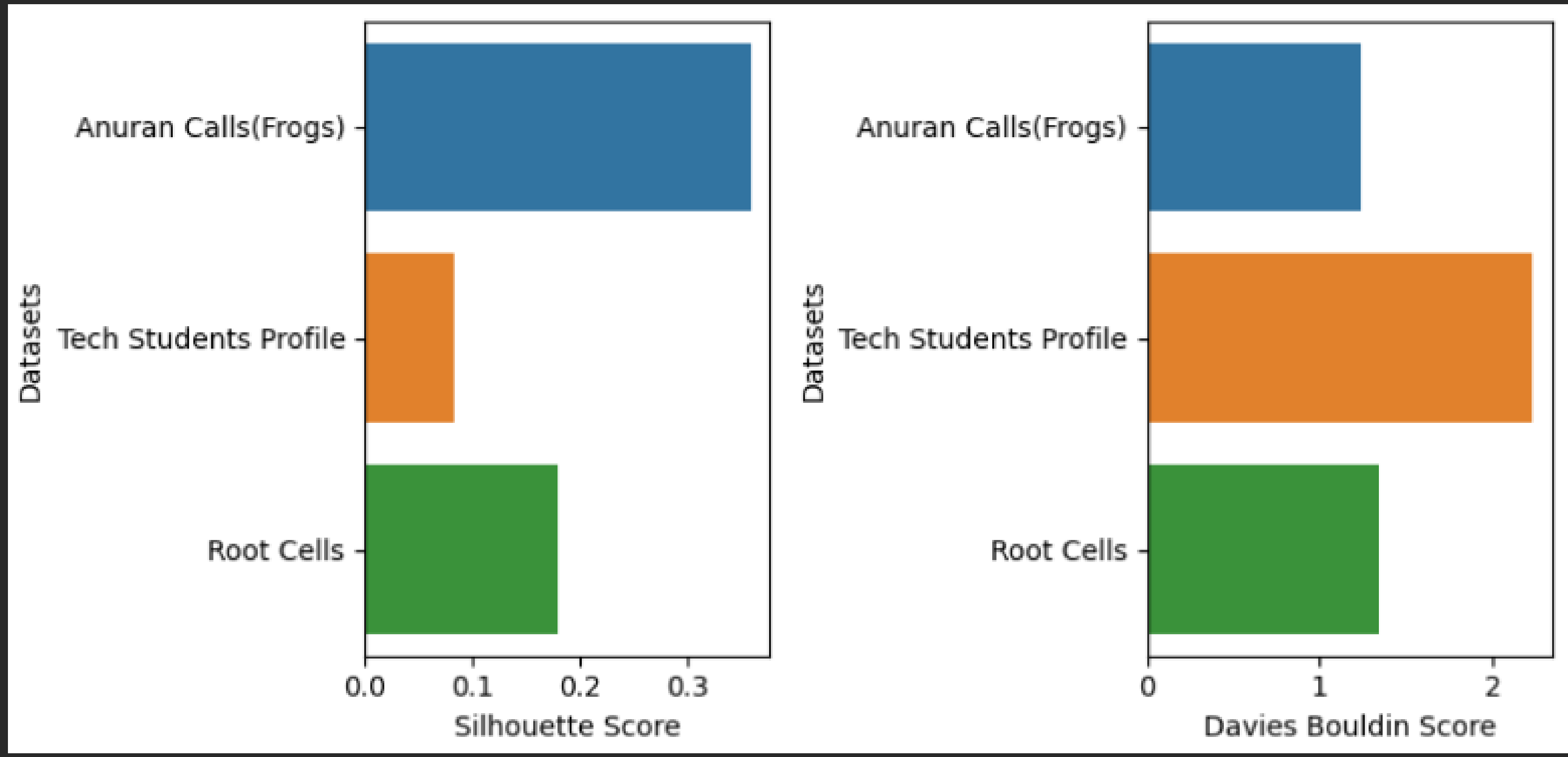
Number of Dimensions: Spectral Clustering

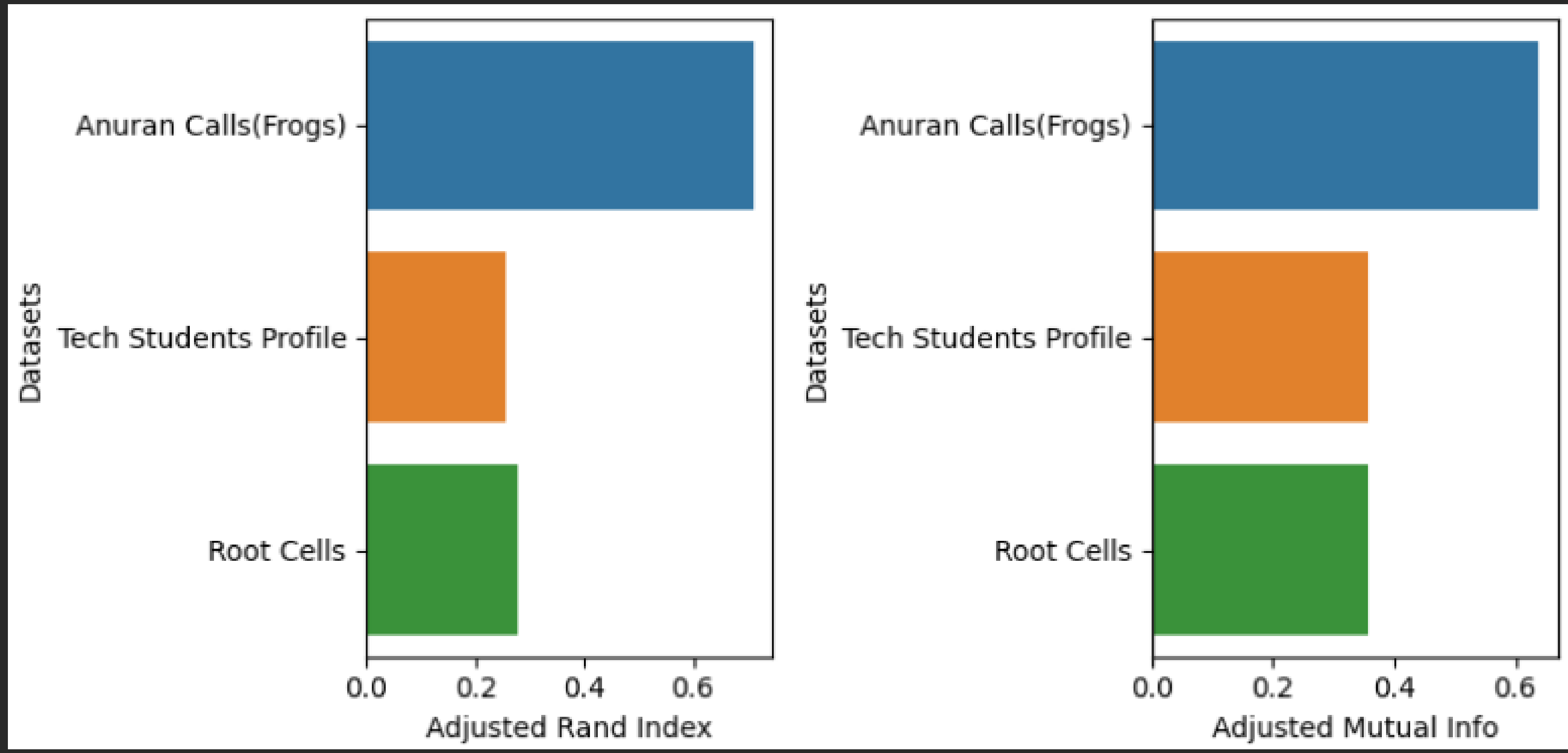
K means Performance



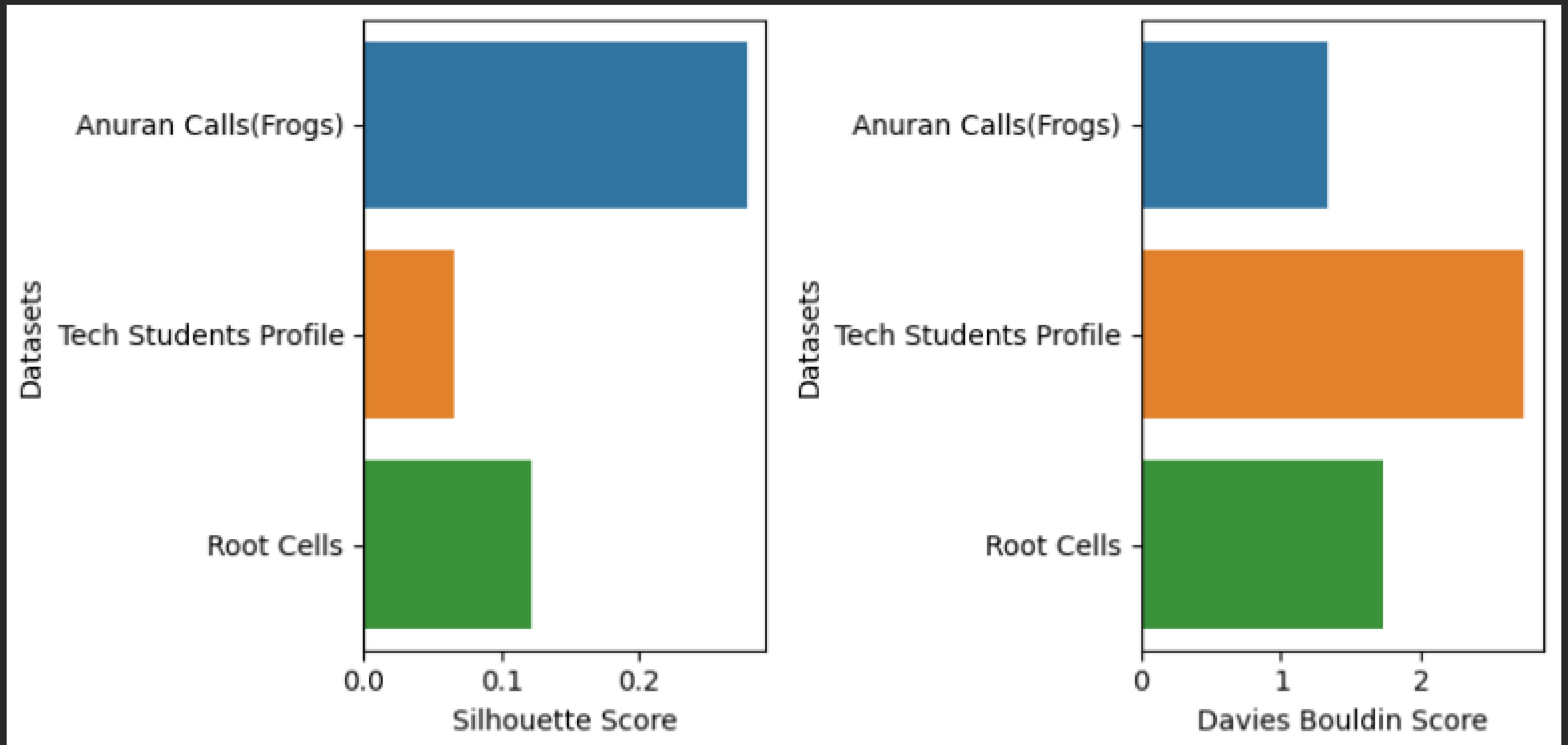


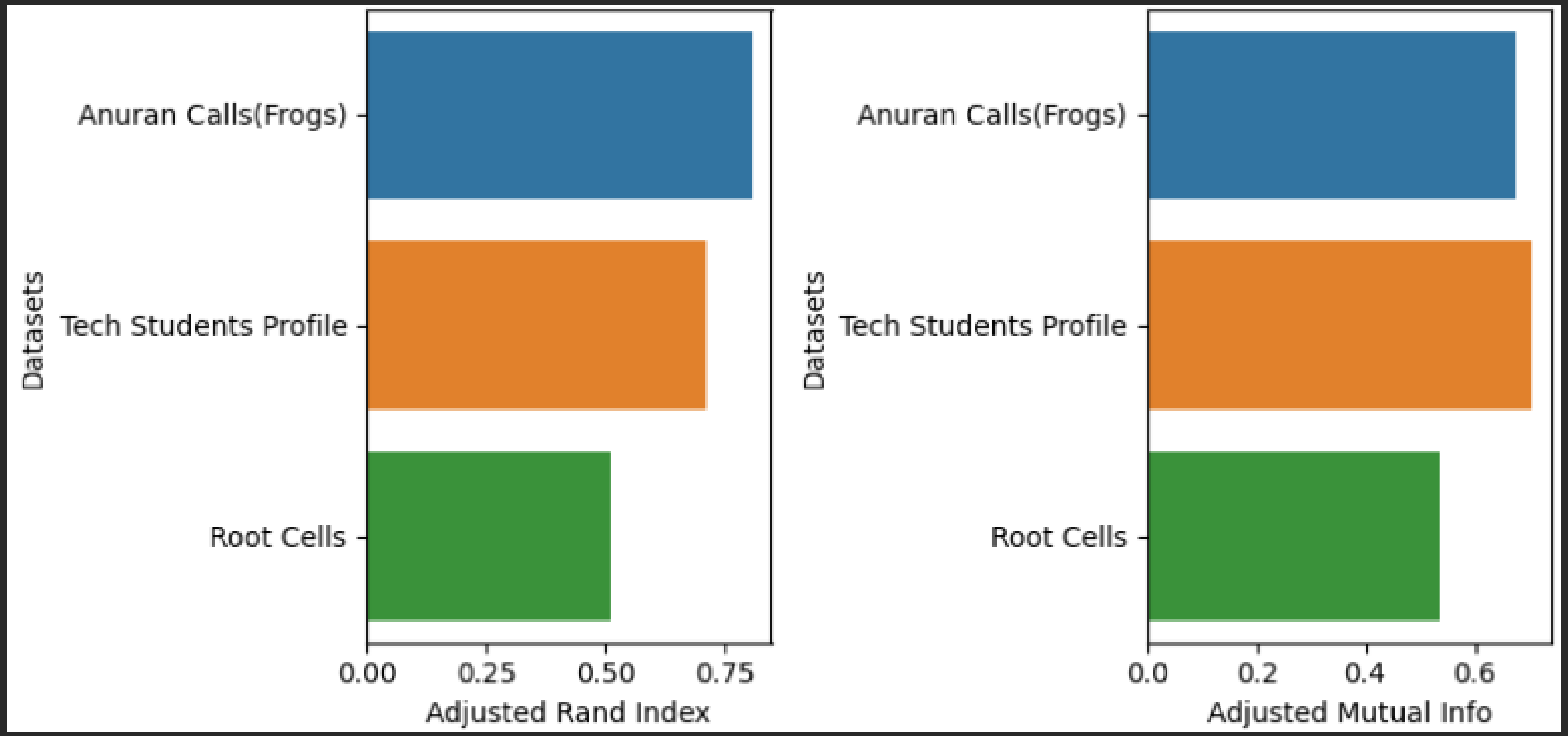
K medoid Performance





Spectral Performance





RESULTS

The best shown results in this case was by the third dataset though other two datasets showed good enough results too.

Thank You