IoT Project 04

TASK 1

1. Hierarchical Clustering:

1.1

For Hierarchical Clustering algorithm, the linkage used was Ward Linkage and Euclidean distance was used for this linkage.

The dendrogram is plotted for the dataset ‘shdeshpa.csv’ and the plot looks as shown below(fig.1).

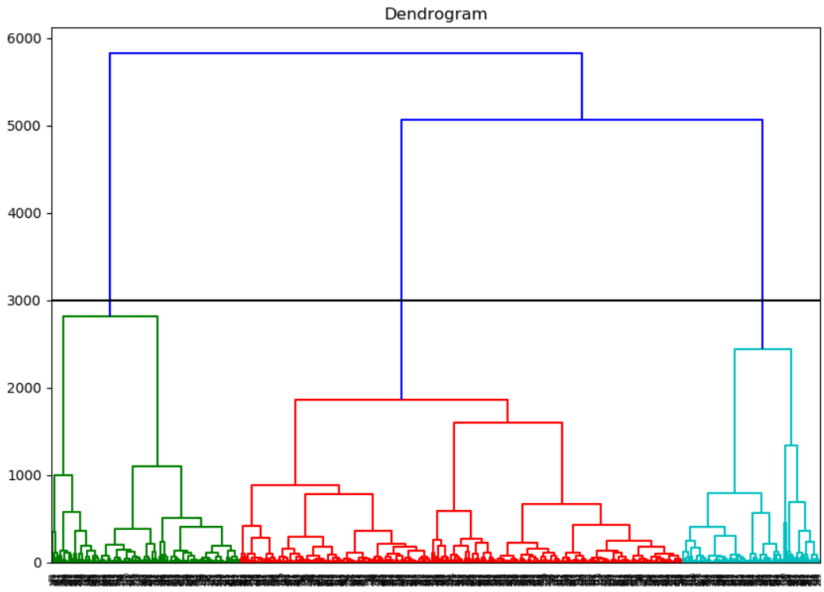

Figure 1: Dendrogram

1.2

From the dendrogram it can be observed that the **number of clusters** are **3**. The dendrogram is cut-off at value ‘3000’ on the y-axis. A cut-off at this point is observed to give optimum number of clusters.

- On the x axis the indices of samples in X are displayed.
- On the y axis the distances (of the 'ward' method in this case) are seen.

1.3 3D Scatter Plot

The 3D scatter plot for the given dataset is plotted using Hierarchical Clustering. It is observed from fig. 2 that the clustering is distinct. However, it can be observed that some points in the purple and green colored clusters have points that are very distant from the main cluster, quite possible noisy data points, are also included in the respective clusters. The fact that noisy data points are included in the clusters can be one of the disadvantages of Hierarchical Clustering.

The clusters were colored differently using the command **cmap = 'rainbow'.**

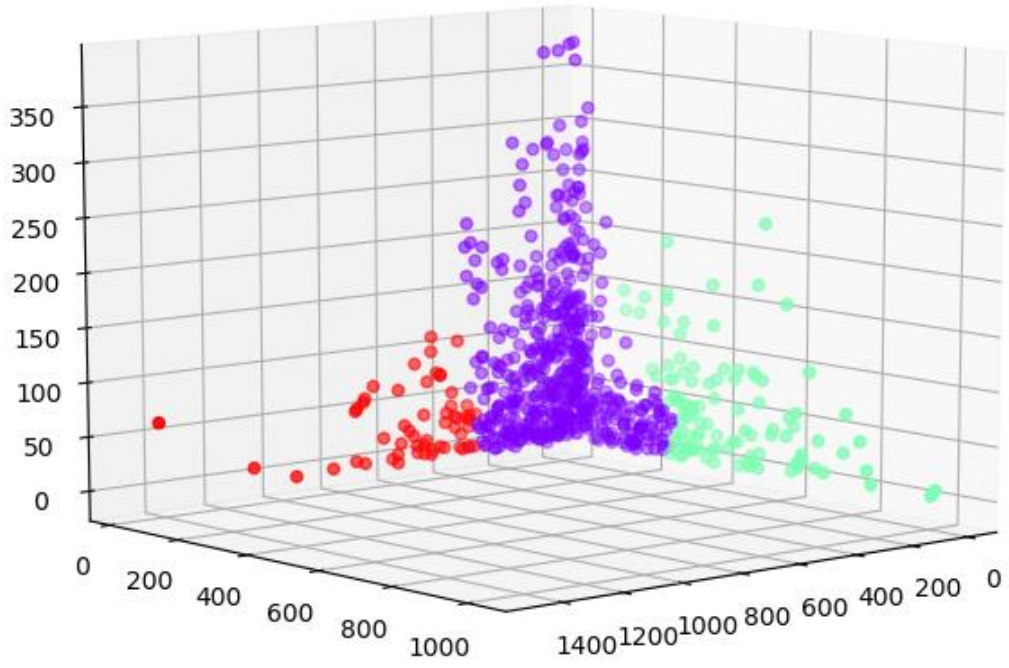

Figure 2: 3D Scatter Plot-Hierarchical Clustering

2. k-means Clustering:

2.1

The k-means algorithm was applied for several values of k. The range specified for values of k was (2,10).

2.2

The elbow method was used to determine the best value of k. The elbow plot is given in fig. 3.

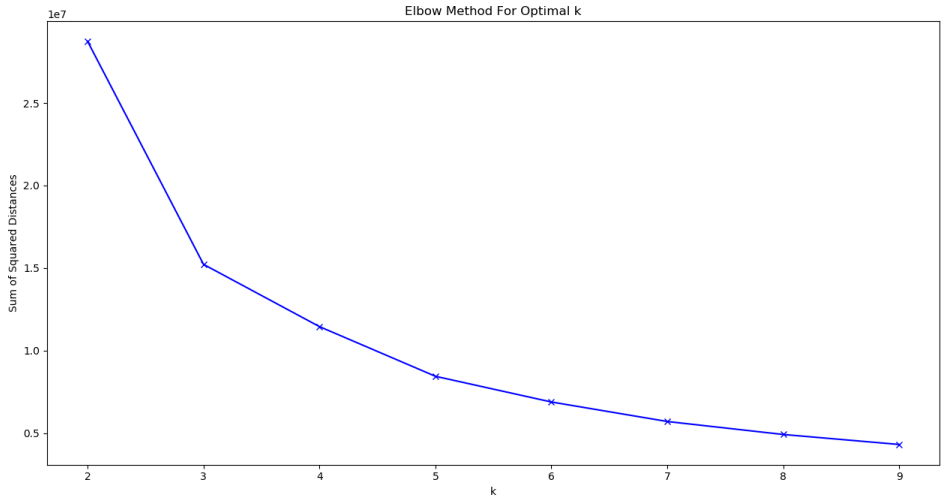

Figure 3: Elbow Plot for k-Means Algorithm

From the elbow plot (fig. 3), it is observed that an elbow is observed at **k=3**. To corroborate the fact that the best value of k is observed at k=3 , the Silhouette scores are calculated.

```
For n_clusters = 2 The average silhouette_score is : 0.4630910844874457
For n_clusters = 3 The average silhouette_score is : 0.5090170560977347
For n_clusters = 4 The average silhouette_score is : 0.44872447451941927
For n_clusters = 5 The average silhouette_score is : 0.4600146838512302
For n_clusters = 6 The average silhouette_score is : 0.42841057697068785
For n_clusters = 7 The average silhouette_score is : 0.423110361448954
For n_clusters = 8 The average silhouette_score is : 0.41666220959365113
For n_clusters = 9 The average silhouette_score is : 0.40057711474880825
```

Figure 4: Silhouette Scores

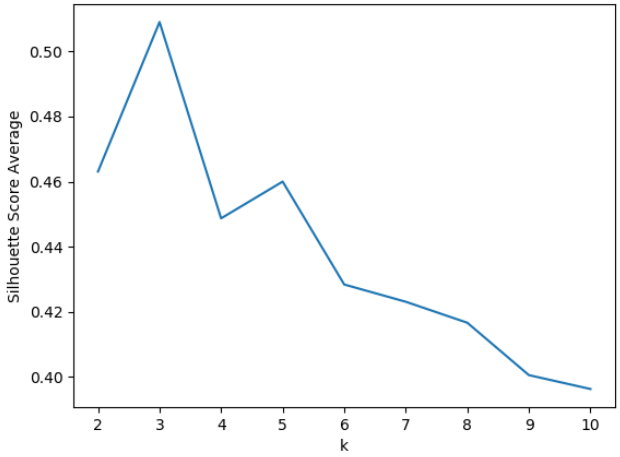

Figure 5: Average Silhouette Score vs k

As seen from Fig. 4 and Fig.5, it is observed that the average silhouette score for k=3 is maximum. Hence, from the elbow method and silhouette scores methods it can be said conclusively that the **best value for k is 3.**

2.3 3D Scatter Plot

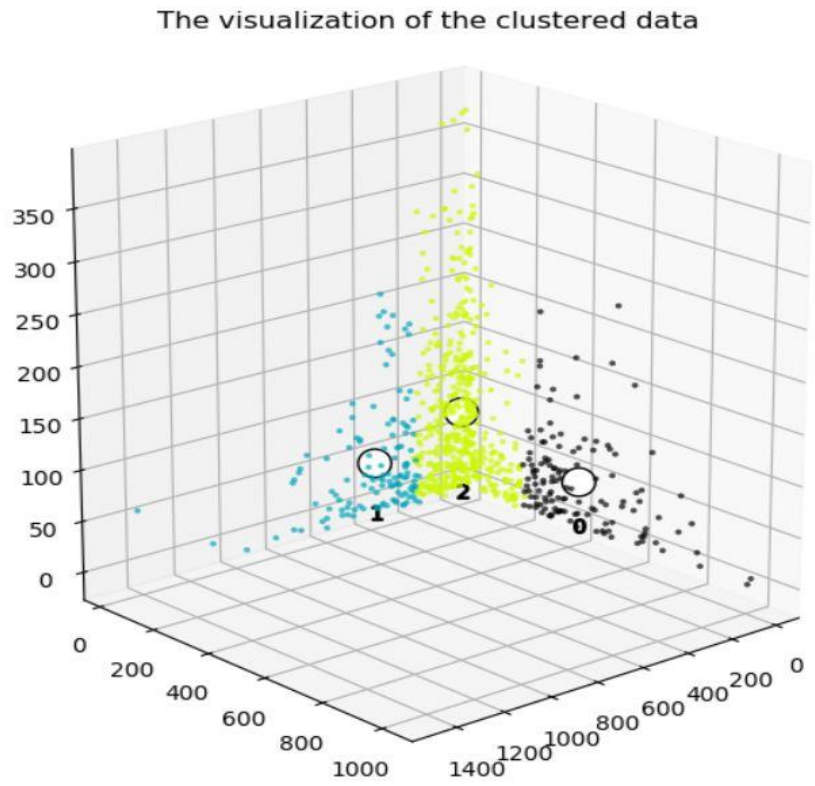

Figure 6: 3D Scatter Plot for k-means

As seen from Fig. 6, each cluster is seen in the 3D scatter plot and is labeled with the cluster id as well.

3. DBSCAN Clustering

3.1, 3.2

The DBSCAN algorithm is applied for MinPts=3,4,5,6 and, epsilon is estimated for this using the elbow method. The radius ϵ is obtained after fixing *MinPts*, by constructing the following graphs (fig.7).

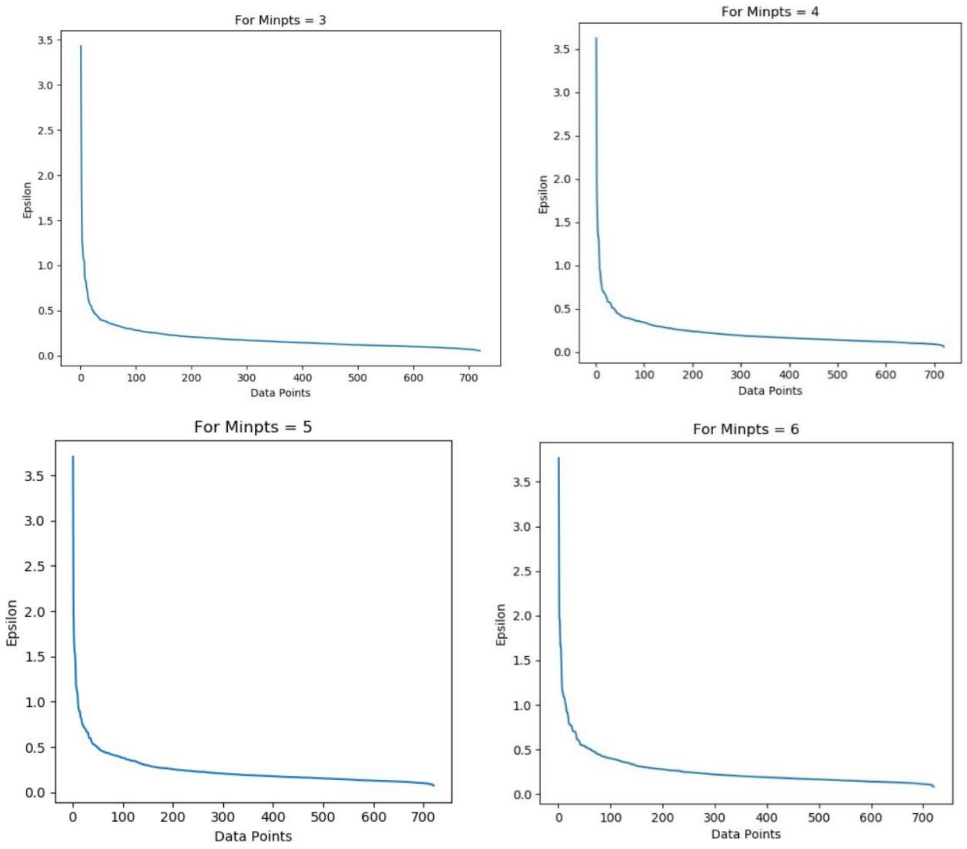

Figure 7: Epsilon Estimation Using Elbow Method

From Fig.7, the best value for epsilon is observed by taking the knee value in each graph and is, Epsilon=0.375 for MinPts=3 and MinPts=4, and epsilon=0.5 for MinPts=5 and Minpts=6.

3.3

By calculating the Silhouette Coefficients(fig.8) for each MinPts, it was observed that the highest value of Silhouette Coefficient is obtained by MinPts = 4.

The same is observed by visualizing the 3D Scatter Plots of MinPts=3,4,5,6. The best clustering is obtained for MinPts=4 at Epsilon=0.375. The 3D scatter plot for MinPts=4 is seen in Fig. 9.

```
Silhouette Coefficient for Minpts=3 is: 0.313

Silhouette Coefficient for Minpts=4 is: 0.367

Silhouette Coefficient for Minpts=5 is: 0.195

Silhouette Coefficient for Minpts=6 is: 0.238
```

Figure 8: Silhouette Coefficients

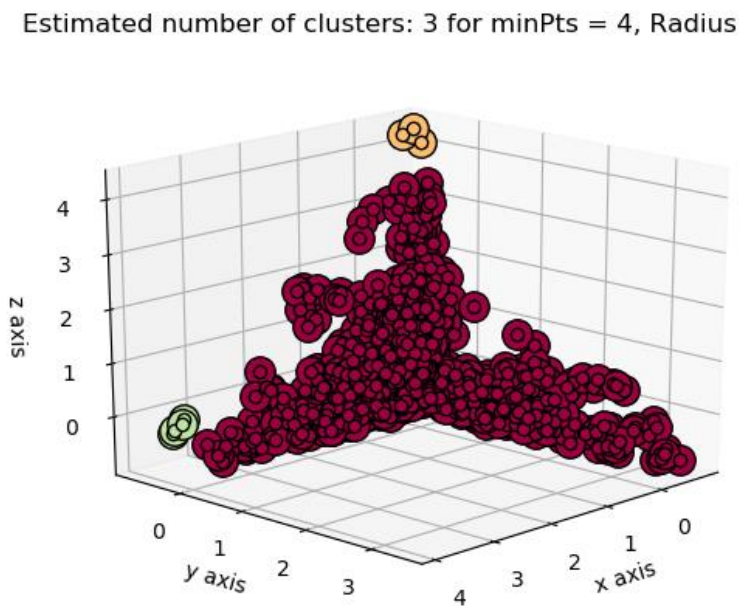

Figure 9: 3D Scatter Plot-DBSCAN Clustering for MinPts=4 and Epsilon=0.375

4.

It is observed that the for the given dataset, all Hierarchical, K-Means and DBSCAN clustering give 3 clusters. But the formation of clusters is quite different in all 3 clustering methods.

Hierarchical Clustering: In this type of clustering, clusters don't have to be predefined. The clusters can be determined from the dendrogram. Hierarchical clustering is sensitive to noisy data hence, it might include noisy datapoints in the cluster. For this dataset, the hierarchical clustering is clearly visible with no overlapping of clusters. Although, the plot includes stray points that might be noisy datapoints. Also, there isn't any distance between the clusters.

K-means Clustering: In K-means clustering, the value of k (number of clusters) must be predefined. The value of k is observed using the elbow method plot. From the 3D scatter plot of k-means clustering it is observed that all clusters are almost of the same size. This is one of the major disadvantages of k-means clustering. Also, there isn't any distance between the clusters.

DBSCAN Clustering: DBSCAN clustering does not require clusters to be specified in advance. This is an advantage as estimating number of clusters for the dataset might be premature. From DBSCAN Clustering's 3D Scatter Plot it is observed that the clusters are

well defined and well-separated unlike the 3D scatter plots of Hierarchical and K-means clustering methods. DBSCAN is robust to outliers as well.

Therefore, it can be said that DBSCAN Clustering Method is the best clustering method for the given dataset.

GAUSSIAN MIXTURE DECOMPOSITION METHOD

1.

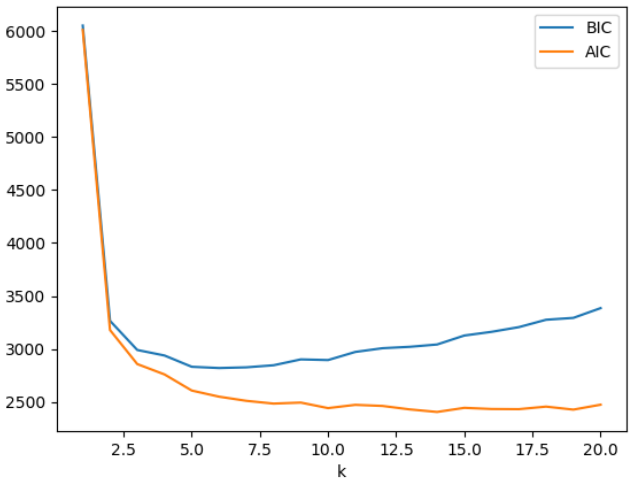

Figure 10: AIC-BIC vs k

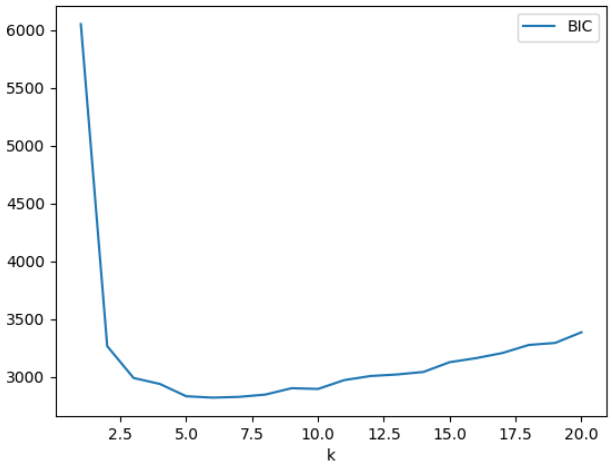

Figure 11: BIC vs k

From Fig.11 it can be said that the best value of **k=5** as this value of k minimizes BIC. The best value of 'k' is the one that minimizes AIC or BIC. Here**, k=5** is considered as it minimizes BIC.

$$BIC = -2\log(L) + \log(n)k.$$

Figure 12: BIC Equation

In Fig. 12, L is the likelihood of data and k is the number of clusters. BIC is minimized at k=5(fig.11). Hence, **k=5.**

2.

The matrix in Fig. 13 is of the format [n_samples=5, n_clusters=k].

```
[[0.148 0.    0.    0.    0.852]
 [0.099 0.    0.    0.    0.901]
 [0.401 0.    0.    0.    0.599]
 [0.04  0.024 0.242 0.141 0.552]
 [0.    0.185 0.    0.815 0.   ]]
```

Figure 13: Probability that a given datapoint belongs to a certain cluster

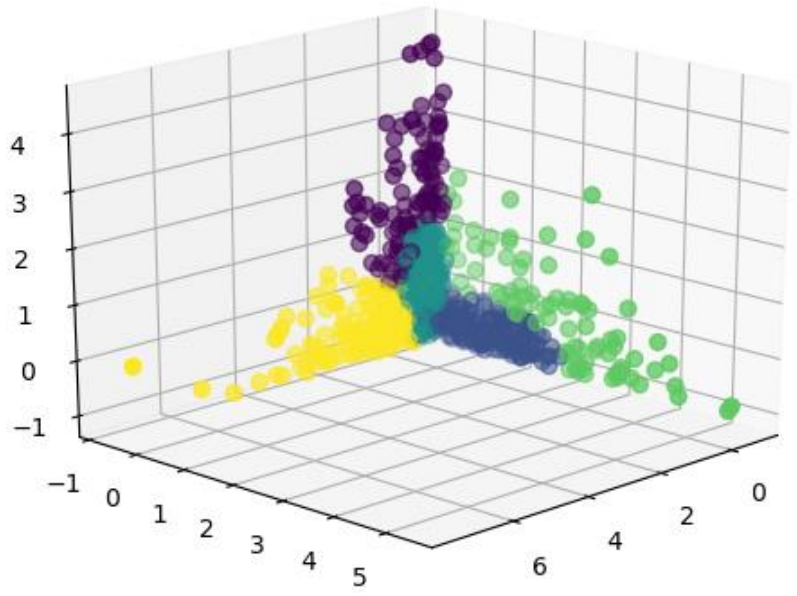

Figure 14: 3D Scatter Plot for k=5

The 3D Scatter plot for a Gaussian Mixture Model is as shown in Fig. 14.

3.

From Gaussian Mixture Models, it is observed that the probability of each datapoint belonging to a certain cluster can be calculated and hence this helps in cluster determination of datapoint. With this probability, it can be determined which distribution the datapoint comes from. For a given datapoint, the cluster for which the probability is maximum, is the cluster to which the datapoint belongs.

From Fig.14 it is observed that there are 5 clusters. They slightly overlap and there is not enough distance between any of the clusters. The clusters don't have to be predefined like K-Means clustering method. However, like hierarchical and k-means they do have noisy datapoints included in the clusters. The clusters aren't well defined and well-spaced like DBSCAN. Hence, for the given dataset, DBSCAN clustering method is the best method for clustering.