Priya Diwakar
200205361

IOT Analytics

Project 4 Clustering

1. Visualize Data

Figure 1 shows the data in a 3D scatter diagram.

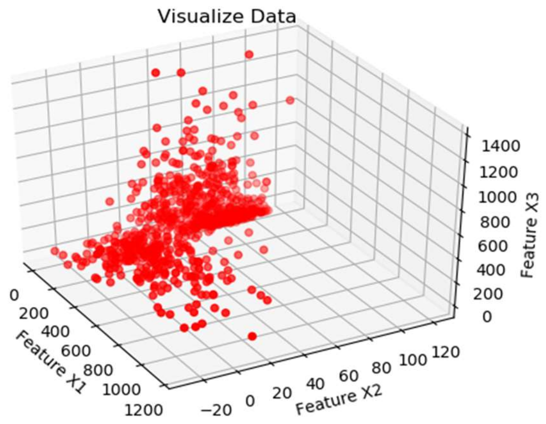

Fig. 1 Data

2. Task 1 Hierarchical Clustering

Complete linkage method is used for distance matrix calculation for the dendrogram. Figure 2 shows the dendrogram. The number of clusters determined from the dendrogram is 3 clusters. Figure 3 shows the 3D scatter diagram for hierarchical clustering.

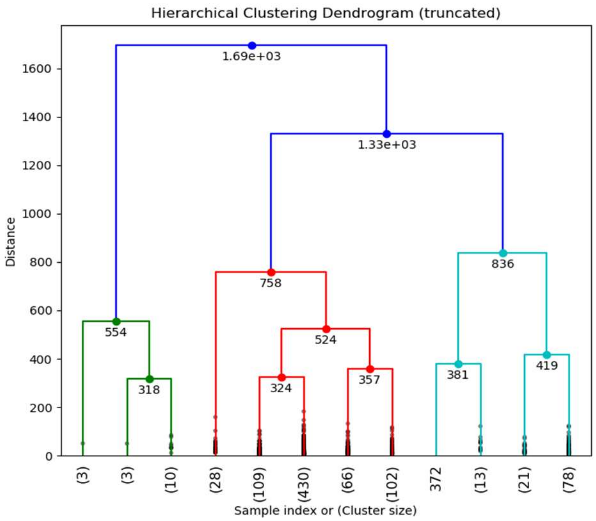

Fig. 2 Dendrogram

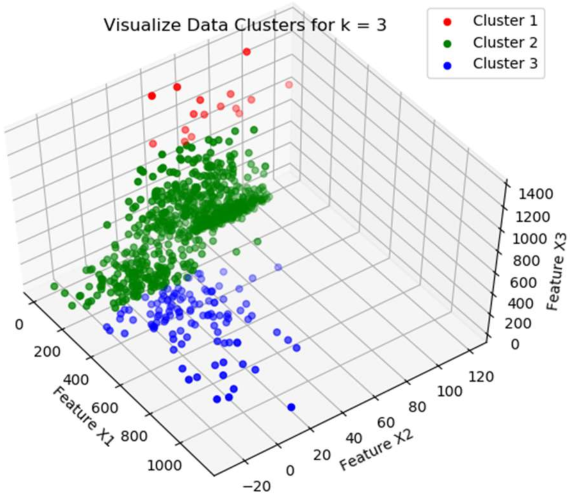

Fig. 3 Hierarchical Clustering

3. Task 2 K means Clustering

For different values of k i.e. the number of clusters, the sum of squared error is calculated. The plot for elbow method is shown in Figure 4. The plot does not show a very clear elbow; hence the silhouette score is also measured. Figure 5 shows the plot of silhouette score versus k. The k for highest silhouette score is found to be 3 and so number of clusters is chosen as 3. Figure 6 shows the 3D scatter diagram for K means clustering.

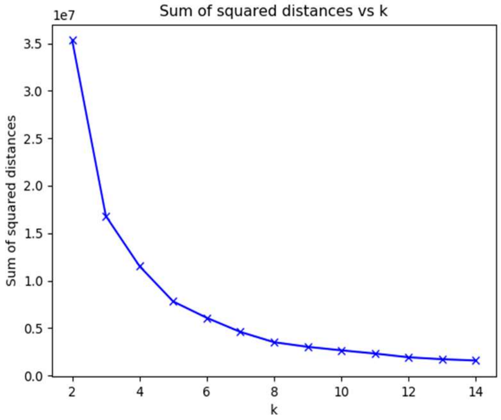

Fig. 4 Elbow Method plot

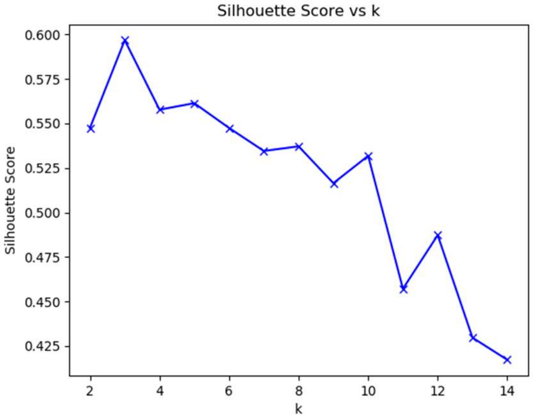

Fig. 5 Silhouette Score plot

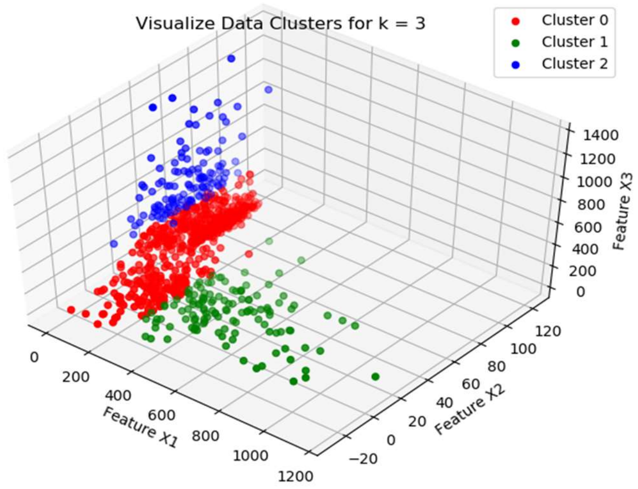

Fig. 6 K Means Clustering

4. Task 3 DBSCAN Clustering

The epsilon value for DBSCAN was obtained by plotting the distance graph. Distances for each point to the closest minimum number of points are obtained. The values are plotted in descending order to determine the knee in the graph. For the dataset given the best clustering was obtained using Minpts = 5 and epsilon ε = 23. The distance graph is shown in Figure 7. Figure 8 shows the 3D scatter diagram for DBSCAN clustering with 4 clusters estimated by the DBSCAN algorithm.

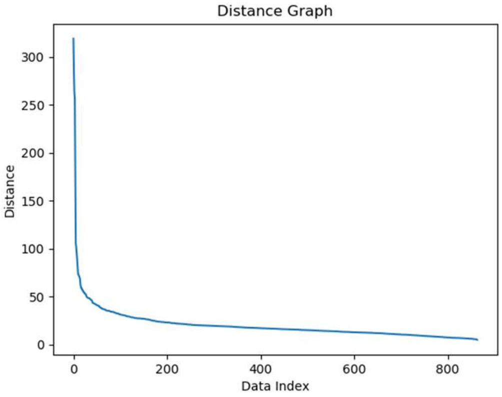

Fig. 7 Distance Graph

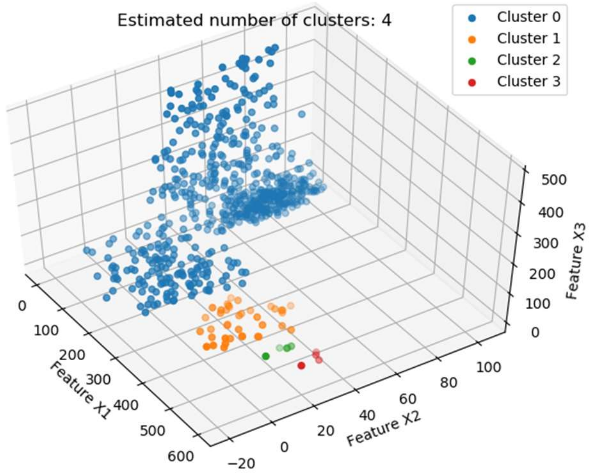

Fig. 8 DBSCAN Clustering

5. Task 4 Compare and discuss results from all three methods

Upon visually inspecting the 3D scatter plots for each of the above method Hierarchical Clustering gives the best results for this dataset. For K means the clustering isn't as good, a few points seem to be misclassified. For DBSCAN the clusters are good but cluster 0 could be 2 clusters upon observing the plot.

6. Extra Credit Task Gaussian Decomposition Method

A. Task 1 Plot the Maximum Likelihood against k

Figure 9 shows the plot of the Maximum Likelihood versus k and it is found that the number of components for the gaussian mixture model equal to 5 gives the highest maximum likelihood.

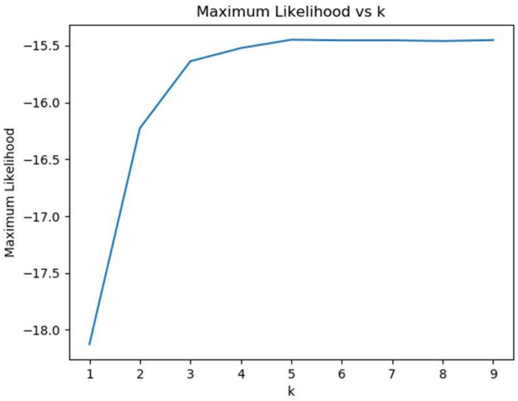

Fig. 9 Maximum Likelihood versus k

B. <u>Task 2 Obtain projections on each plane</u>

The projections for each plane are obtained, i.e. XY, XZ, YZ planes. Figure 10, 11, 12 show the projections for XY, XZ, YZ planes respectively. Figure 13 shows the projections in the 3D graph.

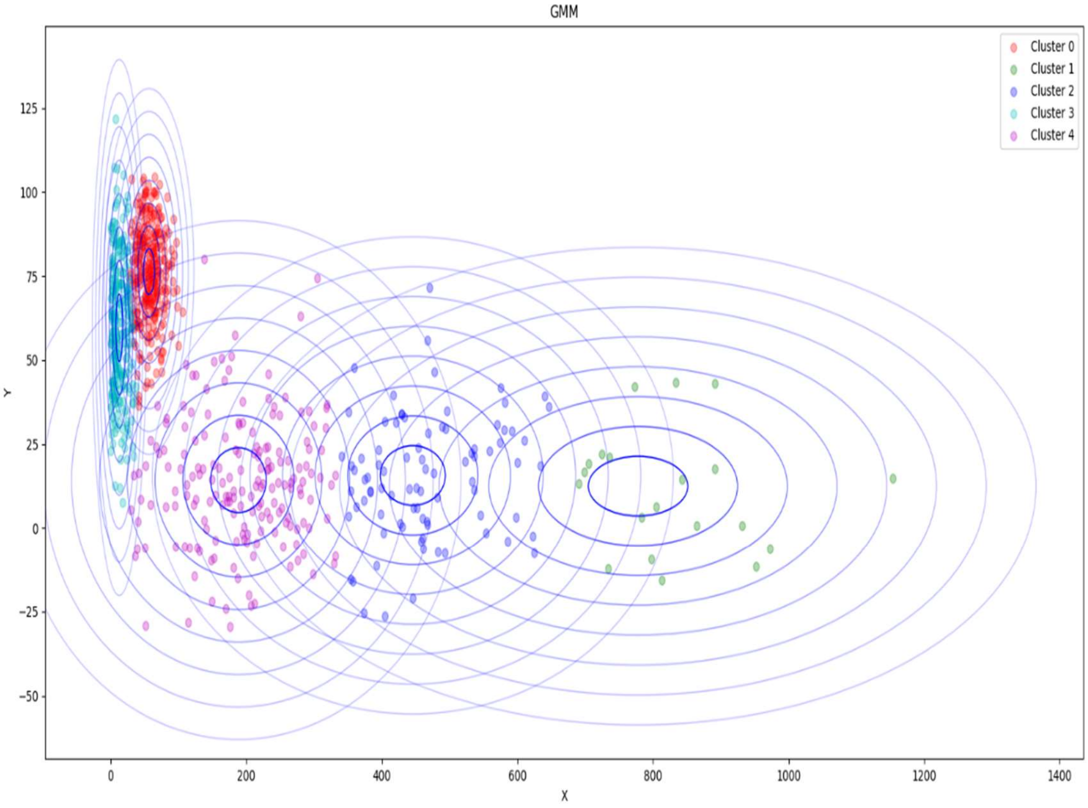

Fig. 10 XY plane projection

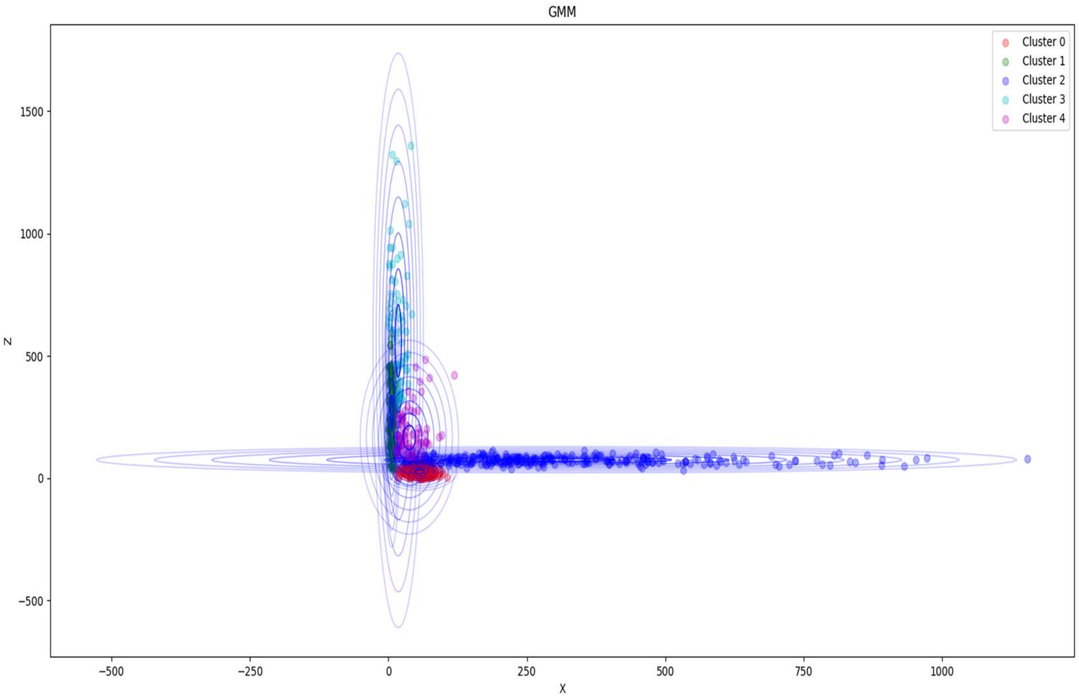

Fig. 11 XZ plane projection

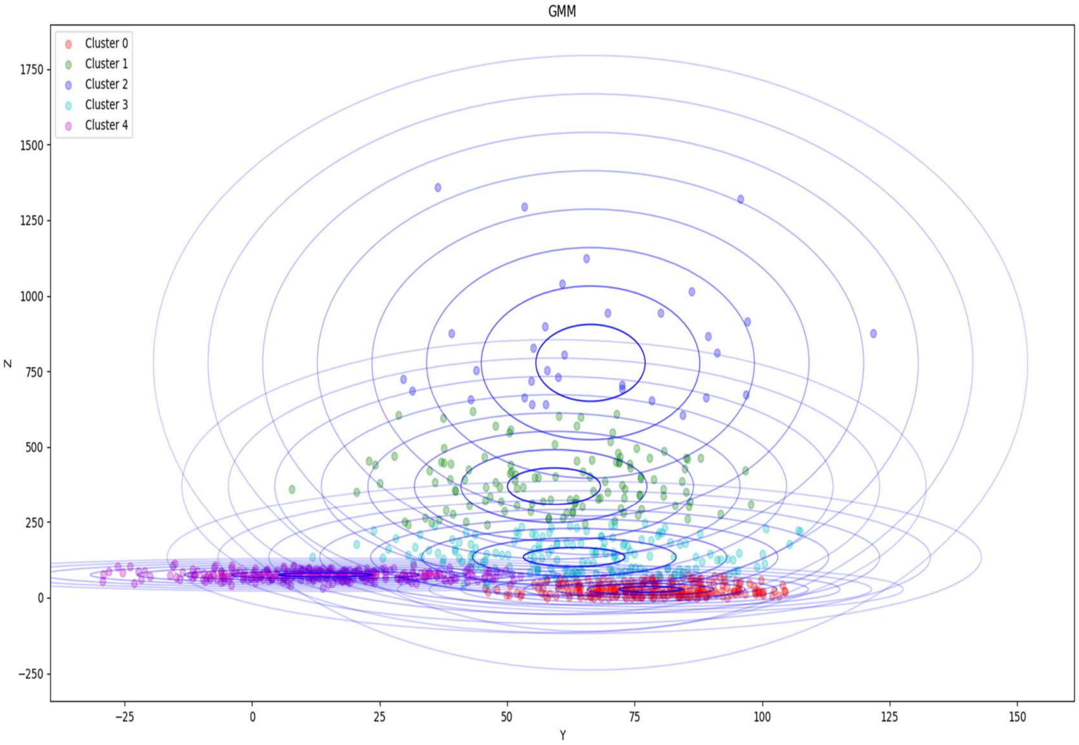

Fig. 12 YZ plane projection

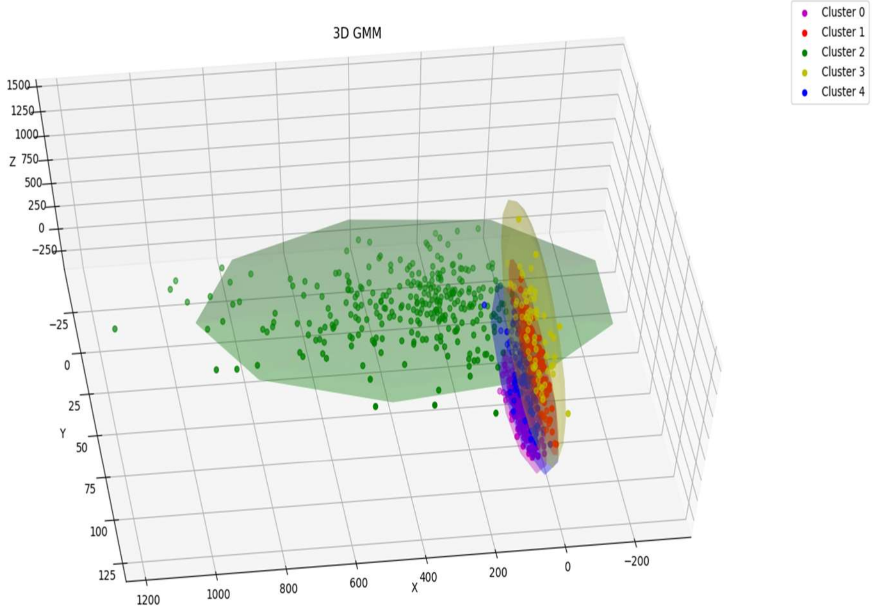

Fig. 13 GMM 3D graph

C. Task 3 Compare results to those obtained from previous methods

The clusters obtained are distinct and data seems to be classified quite well using 5 components for the gaussian mixture model. This is clearly observed from the above plots. The data has been clustered better than the hierarchical clustering.