CS60050: Machine Learning Mini Project 3

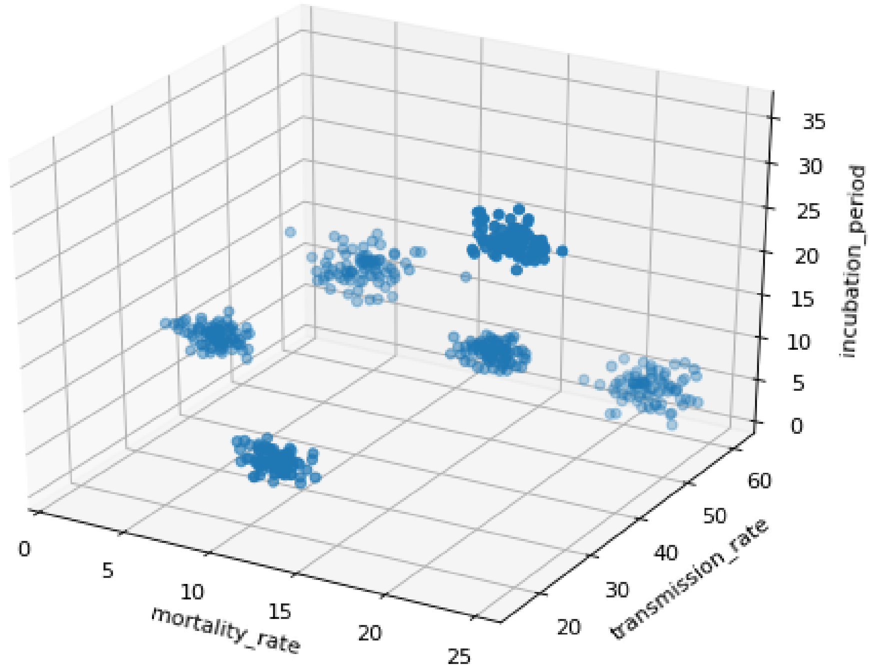
DC-4: Coronavirus Data Clustering using Complete Linkage Hierarchical Clustering Technique

Project Report

Debjoy Saha, 18EC30010

Dataset: Provided dataset had 3 attributes - mortality rate, transmission rate and incubation period belonging to different strains of Corona-Virus.

Dataset Visualisation



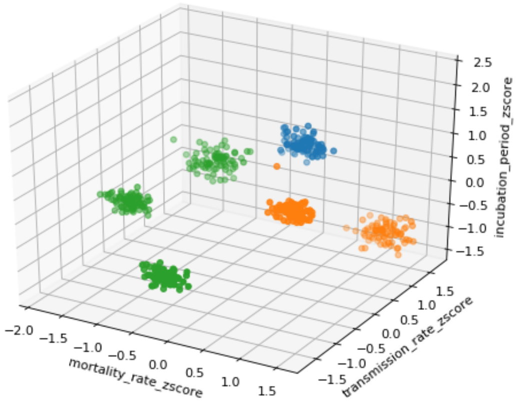
- As apparent, given data had 6 disjoint cluster, each corresponding to a separate corona-virus strain.
- In K-Means clustering, the overall **Best Silhouette score** was obtained as **0.818** for **6 clusters**. However, it varied across runs and depended upon the centroid initialisation for K-means.

Optimal Number of Clusters Obtained = 6

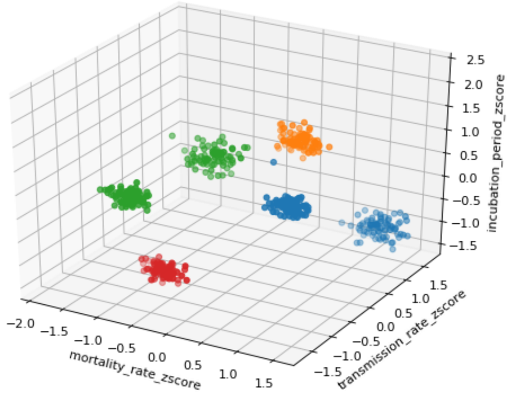
Silhouette Scores obtained for K-means clustering with different K-values -

K-values	Silhouette Score
3	0.568
4	0.670
5	0.714
6	0.818

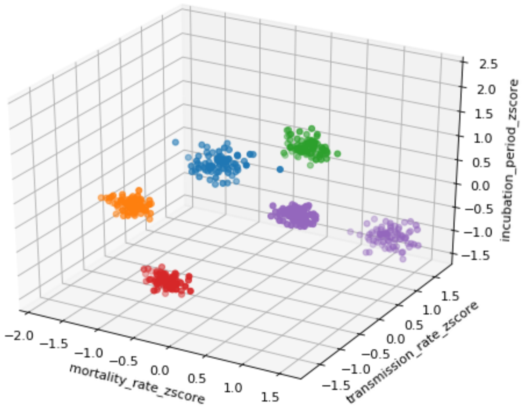
Clusters obtained for different K = 3, 4, 5, 6



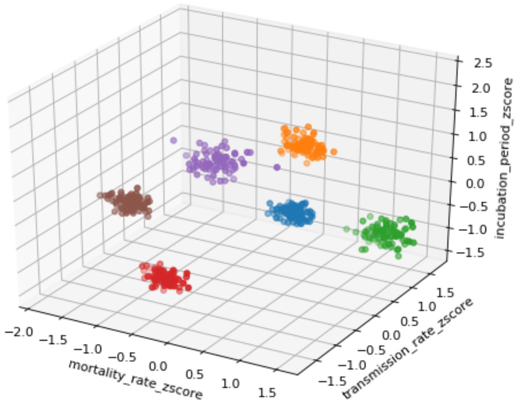
K = 3



K = 4



K = 5



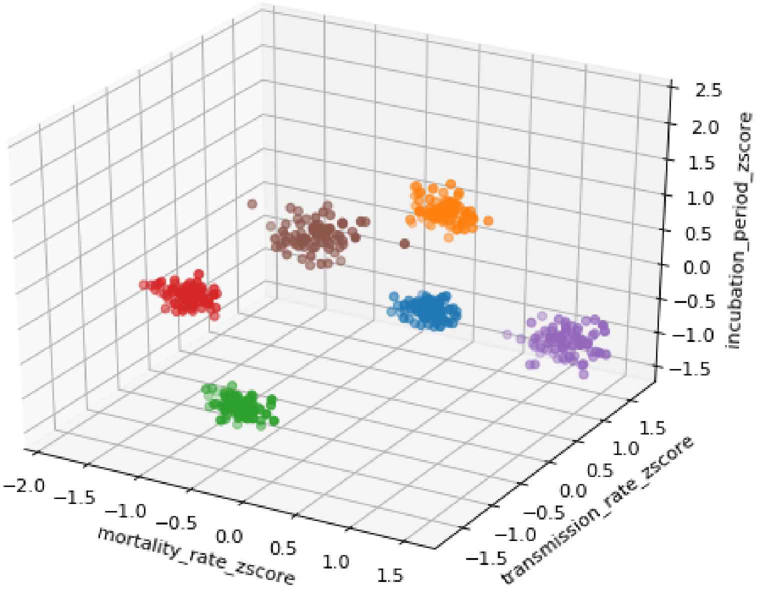
K = 6

- Silhouette Score Obtained for bottom-up Hierarchical Clustering Algorithm using complete linkage strategy with **6 clusters** was obtained as **0.818.**

Cluster Mapping and Jaccard Similarity Score -

K-means cluster id	Agglomerative Cluster Id	Jaccard Similarity
0	4	1.0
1	1	1.0
2	3	1.0
3	0	1.0
4	5	1.0
5	2	1.0

Best Clustering performance Obtained for Agglomerative Clustering with K = 6



- Comparing with clustering performance obtained for K = 6 clusters in K-means classification, we observe an exact match between different clusters. That is the reason we are getting Jaccard Similarity Score of 1.0 for all cluster mappings.

Discussions and Conclusions

Centroid Initialisation: As can be observed from all data visualisation plots, there are 6 disjoint clusters with high inter-cluster separation, for which we are obtaining such a high Silhouette Score Value.

However, a disadvantage is that the clustering performance depends a lot upon the centroid initialisations. Since the clusters are so far apart, if two centroids are initialised in the same cluster, they usually do not shift to different clusters, but rather divide an existing cluster into halves. This results in sub-optimal clustering performance.

One fix is to initialise using **K-means++ algorithm**. In this algorithm the centroid initialisation is performed using a simple rule. The first centroid is initialised at random. All subsequent centroids are sampled from a probability distribution with their probability of selection being directly proportional to the minimum distance from any one (already initialised) centroid. This ensures that the obtained centroids are maximally distant to each other, which increases the probability of their being present in separate clusters.

K-means++ works, but it still doesn't guarantee optimal clustering. The results presented correspond to the best clustering performance obtained across multiple runs.

Silhouette Coefficient: Silhouette coefficient for each data point is defined as S=(b-a)/max(a,b), where a and b are the average L2 distance from data-points belonging to the same cluster and the closest neighbouring cluster respectively. Ideally, b >> a, so S is approximately equal to 1. For determination of closest neighbouring cluster, we used the cluster centroids for K-means, and the final distance matrix for agglomerative clustering. We obtained best average silhouette coefficient of 0.82, thus demonstrating good clustering performance.

Jaccard Similarity Score: The Jaccard similarity score for two clusters is defined as the size of their intersection over the size of their union. Since, the clusters obtained for agglomerative clustering exactly matched with the clusters obtained for K-means clustering, their intersection and union were the same and thus the scores were all obtained as 1.