Assignment - 2

Name: Sai Manoj Kumar Penikalapati
Student ID: 25241477
Class Code: CT5165

Understanding the Dataset:

The modified heart failure dataset contains some clinical measurements for heart-related conditions. Some of the features are Age, Anaemia, Platelets Count, Creatinine Phosphokinase, Diabetes, High Blood Pressure etc.

Since the dataset does not contain **predefined labels**, **unsupervised** learning techniques like **clustering is best suited**. It allows us to group patients based on similarities in their profiles, helping find risk categories.

Among all features, continuous variables such as **age**, **ejection fraction**, **serum creatinine**, **creatinine phosphokinase**, and **platelet count** are expected to influence cluster formation the most because they have high variance and strong correlation with cardiac health.

Data Pre-processing:

Handling Missing Values:

1. **Continuous Variables:** filled using the **median**, which is more robust to outliers and better suited for skewed data. [eg. platelets, sodium, creatinine]
2. **Binary Variables:** filled using **mode**. [eg. anaemia, diabetes, smoking]

As we replace NaN values with some meaningful estimates, the dataset becomes complete and consistent. This helps for better clustering of the dataset as NaN values may lead to bad clustering.

Feature Scaling:

1. **MinMaxScaler:** scales data in range(0, 1), but is sensitive to outliers.
2. **RobustScaler:** uses median and IQR. better suited for medical data and outliers

For this dataset, **Robust Scaler** provided a **higher silhouette score** compared to **MinMaxScaler.** MinMaxSacler is somewhat better compared to StandardScaler.

We are normalising the data in smaller ranges for better distancing in the clusters. For example, platelet count is ranging from 1 to 2 millions. If left unscaled, platelet count would overshadow the influence of other features. By normalising we bring all the features to comparable range, ensuring better cluster formation.

Applying Clustering Algorithms:

KMeans:

- The algorithm randomly initializes k centroids
- Assigns each point to the nearest centroid
- Updates centroid based on mean of assigned points
- Repeats this process until convergence

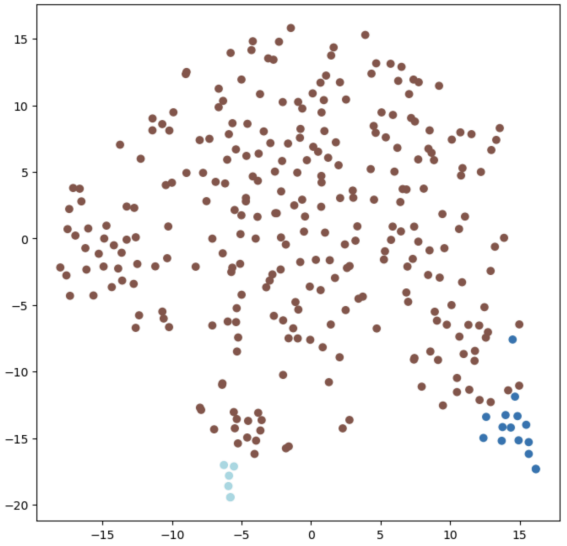
Hierarchical Agglomerative Clustering:

- Starts with each point as its own cluster
- Iteratively merges the two closest clusters
- Produces a dendrogram

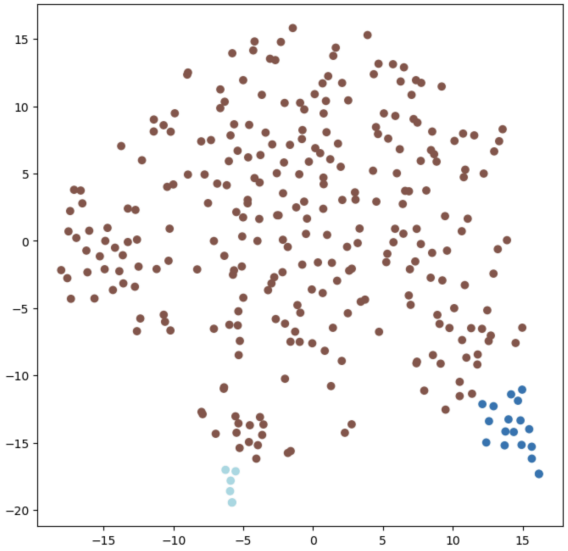
Visualising Clusters:

Here I am using TSNE plot to show how clusters are formed in both the Algorithms

KMEANS



HIERARCHICAL



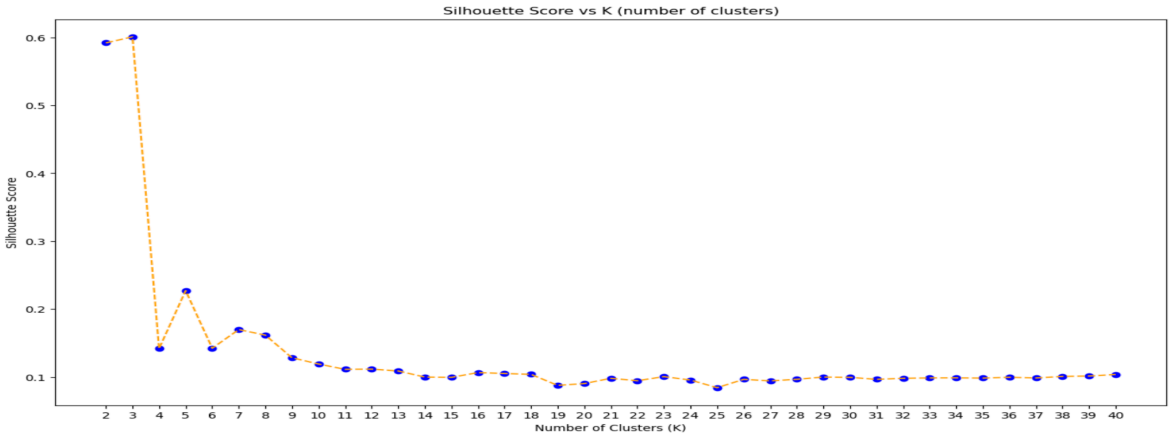
As observed in the images, the clusters produced by both algorithms are largely similar, with only a few minor discrepancies here and there. This consistency supports the conclusion that the clustering effectively identifies distinct types of patients.

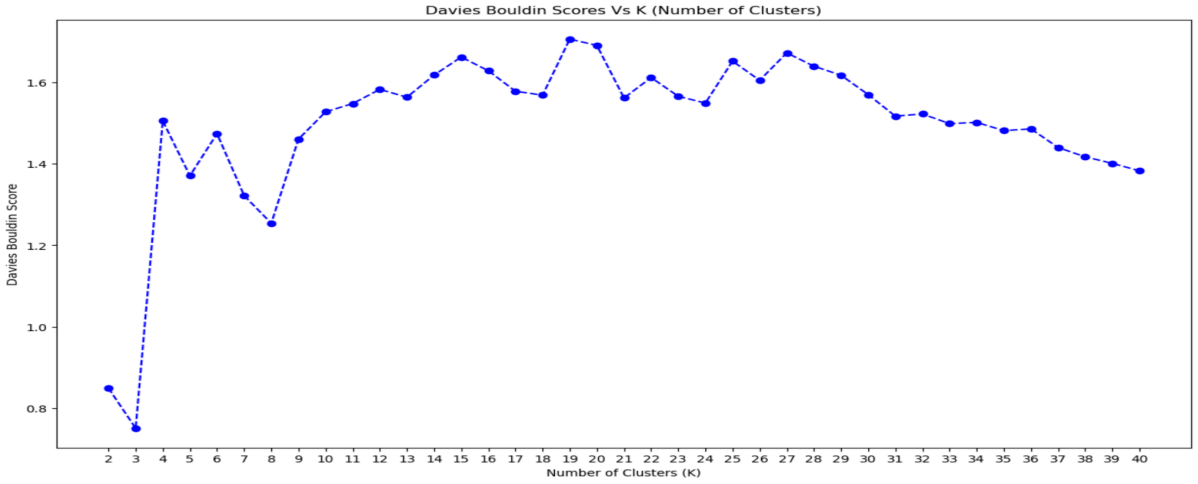
<u>Cluster Evaluation:</u>

For evaluation of clusters formed, I used metrics: **Silhouette Score, Davies-Bouldin Index**

KMeans:

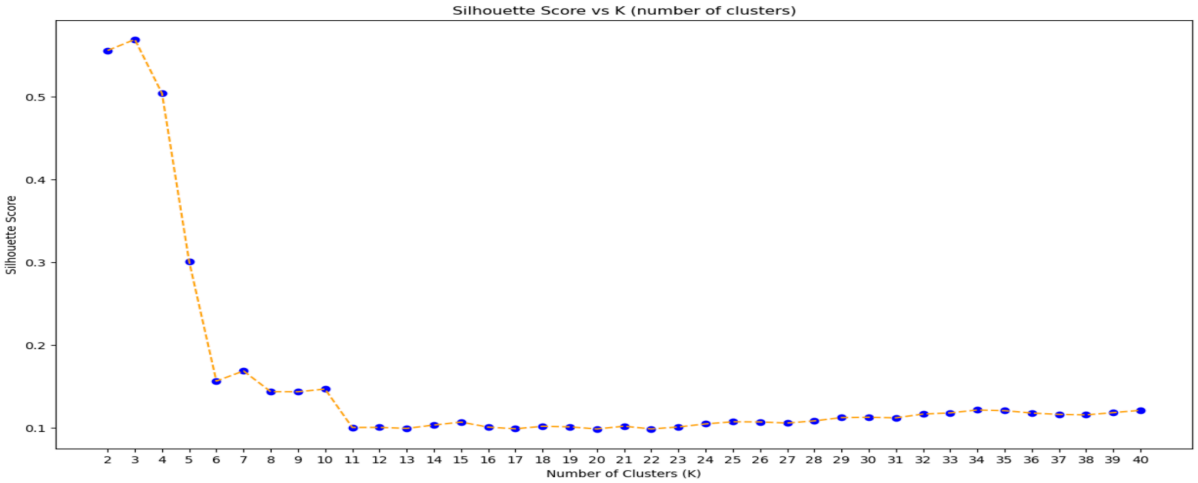
- Silhouette Score and Davies–Bouldin Index were computed for all values of **k** from **2 to 40**.
- The **highest silhouette score (0.6)** was obtained at **k = 3**, indicating well-separated clusters.
- The Davies–Bouldin Index also reached its **minimum at k = 3 (value < 0.8)**, further confirming good cluster separation and compactness.
- Since both metrics agree, **k = 3 is identified as the optimal number of clusters** for this dataset.

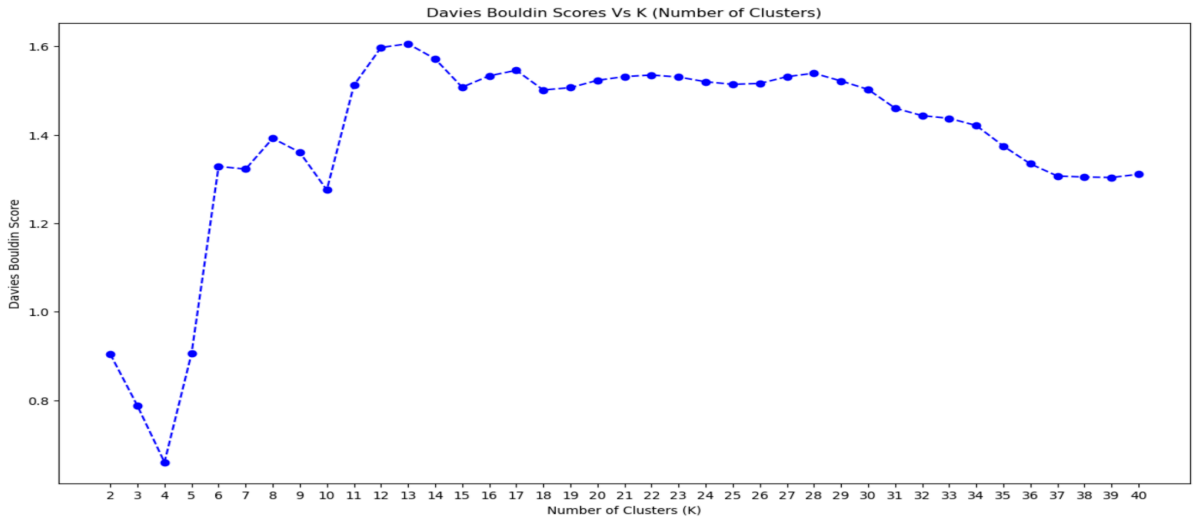




Hierarchical:

- Silhouette Score and Davies–Bouldin Index were computed for all values of **k** from **2 to 40**.
- The **highest silhouette score (0.57)** was obtained at **k = 3**, indicating well-separated clusters.
- The Davies-Bouldin index reached its minimum at k=3,4 but I am considering k=3 because it has a higher silhouette score compared to other k values.
- Therefore, **k = 3** is considered the best clustering configuration for hierarchical clustering on this dataset.





Comparative Analysis:
Both K-Means and Hierarchical Clustering produced very similar and stable cluster structures, indicating a clear underlying pattern in the data. K-Means showed a sharp performance drop after **k = 3**, making the optimal cluster count more evident. In contrast, Hierarchical Clustering showed good performance for both **k = 3 and k = 4**, with a more gradual change from k = 3 to k = 6. Overall, K-Means provided a clearer separation for the best cluster count, while Hierarchical Clustering offered slightly more flexibility but still supported **k = 3** as the most meaningful choice.

Algorithm	**Silhouette Score**	**Davies-Bouldin Index**
KMeans Clustering	0.6007825269810021	0.7507320728991456
Hierarchical Clustering (Agglomerative)	0.5692518089102345	0.7873710254890057

Interpretability and Usefulness:

The clusters are highly interpretable, as they group patients based on meaningful clinical features such as age, ejection fraction, serum creatinine, and platelets. These clusters reflect different levels of cardiac risk—typically **separating low-risk, moderate-risk, and high-risk** patient groups. This provides useful insights into how combinations of medical factors relate to overall heart health.

In real-world healthcare settings, clustering can help identify high-risk patients early, prioritize those needing urgent monitoring, and support personalized treatment planning. Hospitals can use such clustering models for risk stratification, resource allocation, and improving preventive care strategies.