Heart Disease Clustering

I. Describe the data

Doctors often analyze past cases to enhance their understanding of optimal treatment approaches for their patients. When a patient exhibits similar health history or symptoms to a previous case, they may benefit from undergoing a similar treatment regimen.
This challenge aims to explore the possibility of grouping patients based on common characteristics using unsupervised learning techniques. Specifically, we will investigate the use of k-means algorithm. By employing this method, we can identify clusters of patients with similar attributes, facilitating targeted treatment strategies.

II. Objective

My focus is on examining anonymized patients who have received a diagnosis of heart disease. By identifying patients who share similar characteristics, we can gain insights into the effectiveness of specific treatments. This information would be valuable for doctors, as they could learn from the outcomes of patients similar to those they are treating. The dataset originates from the V.A. Medical Center in Long Beach, California.

III. Procedures

In this project, I use three different clustering models to cluster the data: KMeans, DBScan, and AgglomerativeClustering. And then I evaluate each unsupervised learning model's performance using silhouette score.

First, ColumnTransformer was used to preprocess the dataset: Imputing and standardizing. After that 29 columns were created after preprocessing. Thus, PCA was used on it in order to denoise, remove redundancy, and improve clustering performance. And then to select the appropriate number of clusters for KMeans a 'for loop' was used to iterate the k value from 2 to 10 and best k value was selected based on silhouette score. However, silhouette score of each k value applied on KMeans was extremely weak, ranging from 0.161 to 1.190 at most, suggesting that KMeans is not suitable for this dataset.

Second, DBScan was used. Same as KMeans, the dataset was passed through a pipeline: preprocessing, PCA, and then the DBScan class. The performance was evaluated using the silhouette score. The scores were significantly improved compared to KMeans: 0.297 to 0.572. EPS of 0.60 and the number of clusters of 2 obtained the 0.572 of score.

Third, AgglomerativeClustering was used. But it did not give any improvement in score: only ranging from 0.146 to 0.376.

IV. Key findings

It has been found that the DBScan model with the eps value of 0.60 and clusters of 2 are most suitable for this dataset.

V. Reflections

Possible reasons that KMeans and AgglomerativeClustering struggled at this dataset whereas DBScan excelled are:

- Clusters are not equally sized or shaped.
- Data contains outliers: It is important to check outliers before proceeding with the task.
- Clusters are not linearly separable.
- Silhouette Score is low because many points are closer to neighboring clusters than their own centroid: it is a sign KMeans forcing data into clusters it does not belong to.
- DBScan does not assume any cluster shape: it can identify clusters of arbitrary shapes.
- AgglomerativeClustering tends to produce hierarchical partitions that do not always match the true data structure.
- DBScan labels outliers as -1, effectively ignoring them in cluster formation, which leads to cleaner cluster separation.
- Silhoutette Score is sensitive to cluster quality.

VI. Future plans

Since the problems with KMeans and AgglomerativeClustering are related to outliers, cleaning outliers is an utmost importance prior to fitting with distance-based unsupervised learning models. Outliers can be easily visualized using boxplot and pinpoint the exact index using IsolationForest.