Prediction of heart disease using K-clustering Algorithm

Unsupervised machine learning is a branch of data analytics that is used for prediction of heart disease. Unsupervised machine learning techniques include K-means clustering. Unsupervised algorithms often forecast the intended result without reference to any values. The K-means clustering technique clusters the data so that there is a high intra-class similarity and a low inter-class similarity. The sum of squares distance from the cluster centroid is reduced by this algorithm. The programme creates k clusters with a centroid out of the data. K-means iteratively locates the centre that minimises the separation between the cluster's individual points and its centre. The k-means clustering technique is demonstrated in the following flow chart.

This section presents the findings of the data analysis conducted to find the necessary hidden patterns for forecasting cardiac illnesses. Age, the type of chest pain, blood pressure, blood sugar level, resting ECG, heart rate, the four different types of chest pain, and exercise-induced angina are the variables here taken into account to predict heart disease. Pre-processing the heart disease dataset efficiently involves removing irrelevant records and assigning values to tuples that are missing. The K-means technique is then used to put together the pre-processed heart disease data set. In this article, four different types of heart diseases-asymptomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain-are explored.

K-means Clustering

The K-means clustering algorithm was chosen for the web dataset due to its scalability, efficiency, ability to construct populations of equal size, and simplicity. In order to classify groups of data points, the K-means algorithm requires a minimum total of squares. The dataset in question has 209 observations for 7 variables. The steps listed below are used to calculate the cluster's initial centre.

1) Detect K random clusters
2) Find the important clusters iteratively
3) If the distance between the observation and its nearest cluster centre is greater than the distance between the other closest cluster centres, the observation is substituted with the nearest centre by computing the Euclidean distance between the cluster and the observation.
4) The total of squares within a cluster is determined as follows:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where S_k is the set of observations in the *k*th cluster and $\bar{x}_{kj}$ is the *j*th variable of the cluster center for the *k*th cluster.

The Final Cluster Centres are reached when the difference between the sum of the squares in two successive iterations is at a minimum.

Age, maximal heart rate, the nature of the chest pain, and the condition are all factors taken into consideration when predicting heart disease. The findings are addressed individually after taking into account four different forms of chest discomfort.