****

**20CSIS01H**

**Data Mining & Warehousing**

**Group Number 1**

**phase 1**

|  |  |  |
| --- | --- | --- |
| Mostafa Ebrahim Negm | SE | 153095 |
| Maryam Ashraf Diaa | SE | 162417 |
| Hussien fawzy | IS | 158267 |

**Dataset description:**

The dataset contains 14 attributes published by the Cleveland database. The data set refers to heart disease in patients. Where takes patient info and valued from 0 that means no presence to 4.

The attributes of the dataset contain:

* Patient Age as age
* Patient gender as sex
* Patient chest pain type as cp
* Patient resting blood pressure as tresrbps
* Patient serum cholestoral as chol
* Patient fasting blood sugar as fbs
* Patient resting electrocardiographic results as restecg
* Patient maximum heart rate achieved as thalach
* Patient exercise induced angina as exang
* Patient oldpeak as oldpeak
* Patient slope of the peak as slope
* Patient number of major colored by flourosopy as ca
* Patient thal as thal
* And target

**Data preprocessing:**

The data has no empty or missing values. It was already preprocessed. But all is done is normalizing the dataset

**K-Means Algorithm:**

K-Means is used to divide data objects into clusters it works by measuring the distance between each point and group them together in one cluster the distance is measured by Manhattan distance.

**Hieratical Algorithm:**

is an algorithm that adds data in hierarchical cluster model. By assigning data points to cluster of their own. And then merge the closest clusters into the same cluster. Which at the end only one cluster remains

**Comparison:** Hieratical clustering cannot handle bigdata as the time complexity is quadratic O(n^2), but the K-Means can handle big data as the time complexity is linear O(n).

**K-Means clustering:**

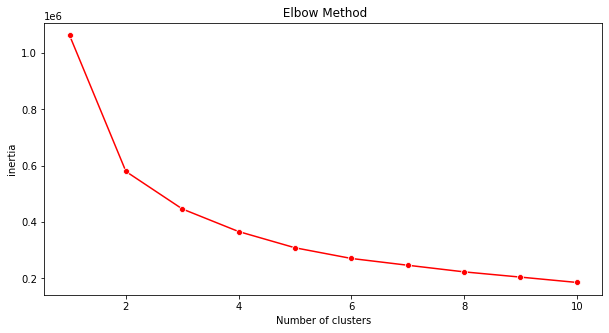
K-Means uses steps to be able to visualize data by doing number of steps as following:

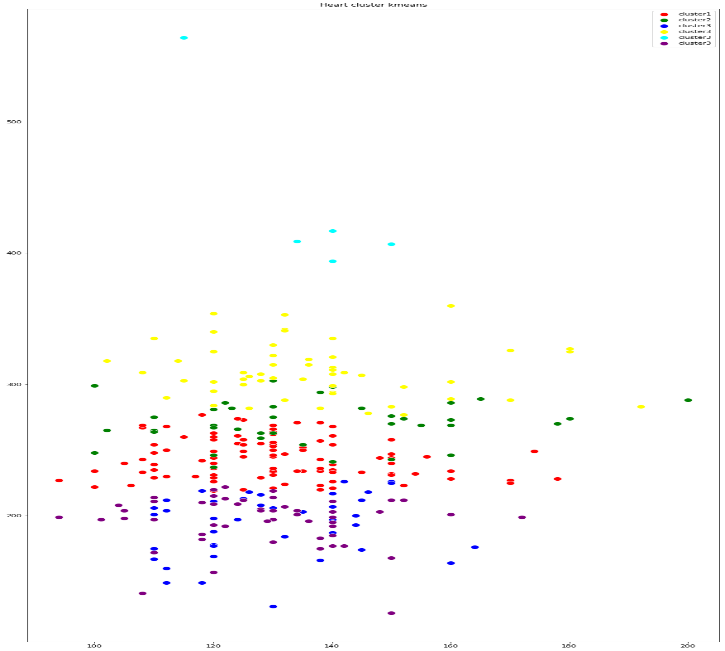
1. specifying number of clusters
2. randomly assign each data point to cluster
3. compute the cluster centroids.
4. Reassigning each point to the closest cluster
5. Re-computer cluster centroids
6. Then repeating steps 4 and 5 until there is no improvements into clusters

The K-Means cluster accuracy alone can be 78% but it can be enhanced by better model to give accuracy of 95%.

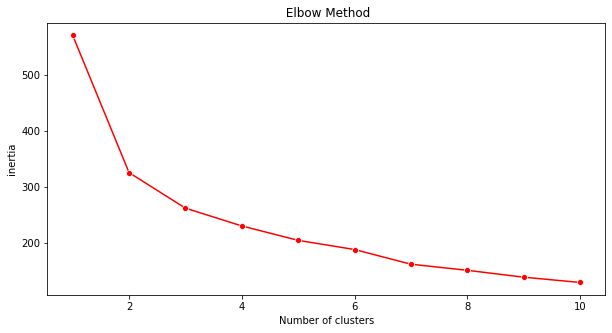
**First code was by takin 3,4 and 7 and by applying the elbow method the optimal clusters was 6 so the visualizing is:**

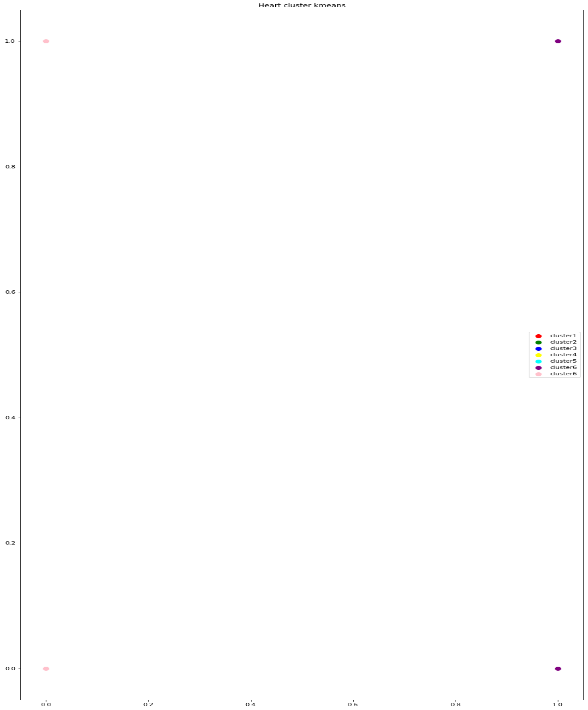
**\*hence: elbow method is method that gets the number of the best clusters after the points of the inertia starts to decrease in linear way**

****

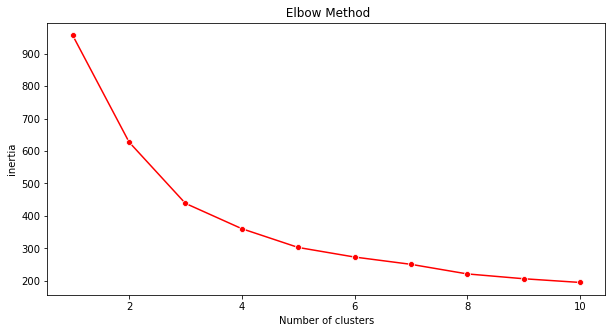
****

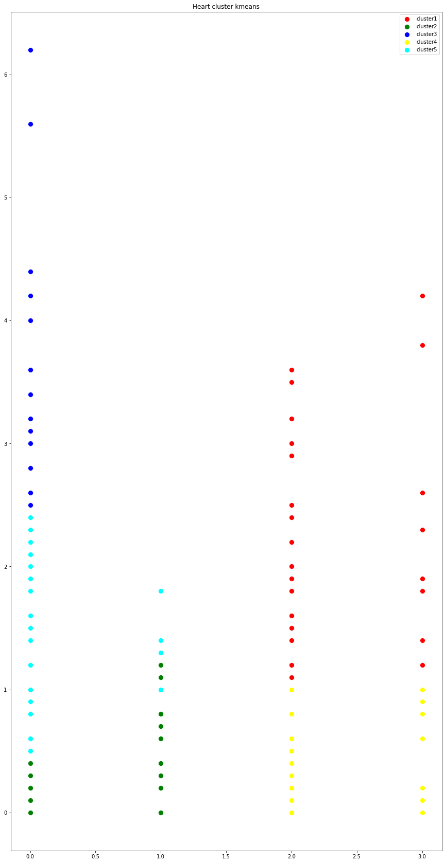
**second code was by takin 1,5,6,8 and 11 and by applying the elbow method the optimal clusters was 7 so the visualizing is:**



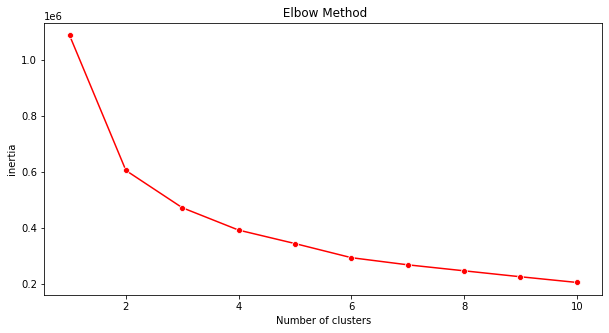


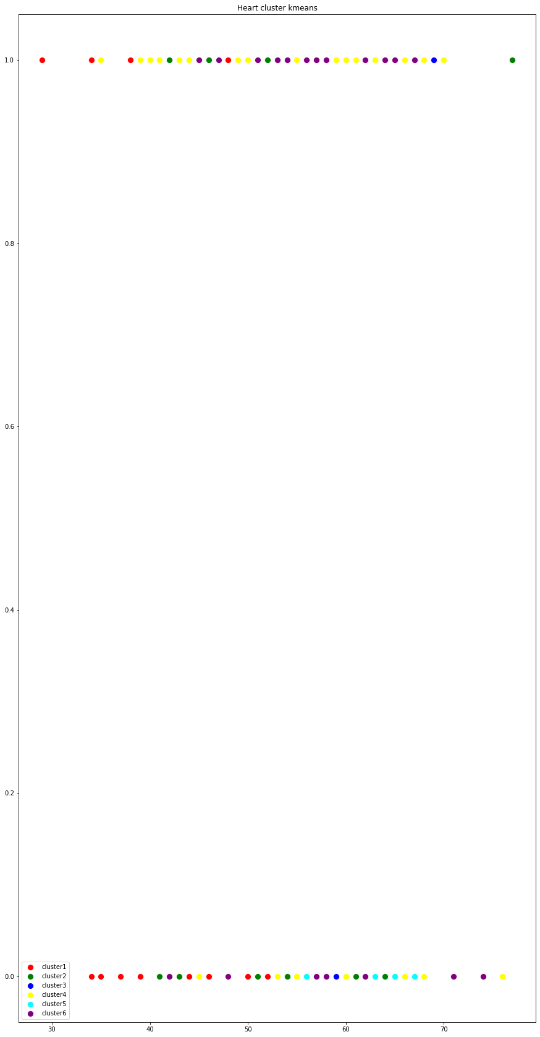
**Third code was by takin 2,9,10 and 12 and by applying the elbow method the optimal clusters was 5 so the visualizing is:**

****

****

**The last code was by takin all the columns and by applying the elbow method the optimal clusters was 7 so the visualizing is:**

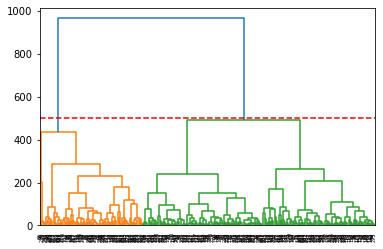
****

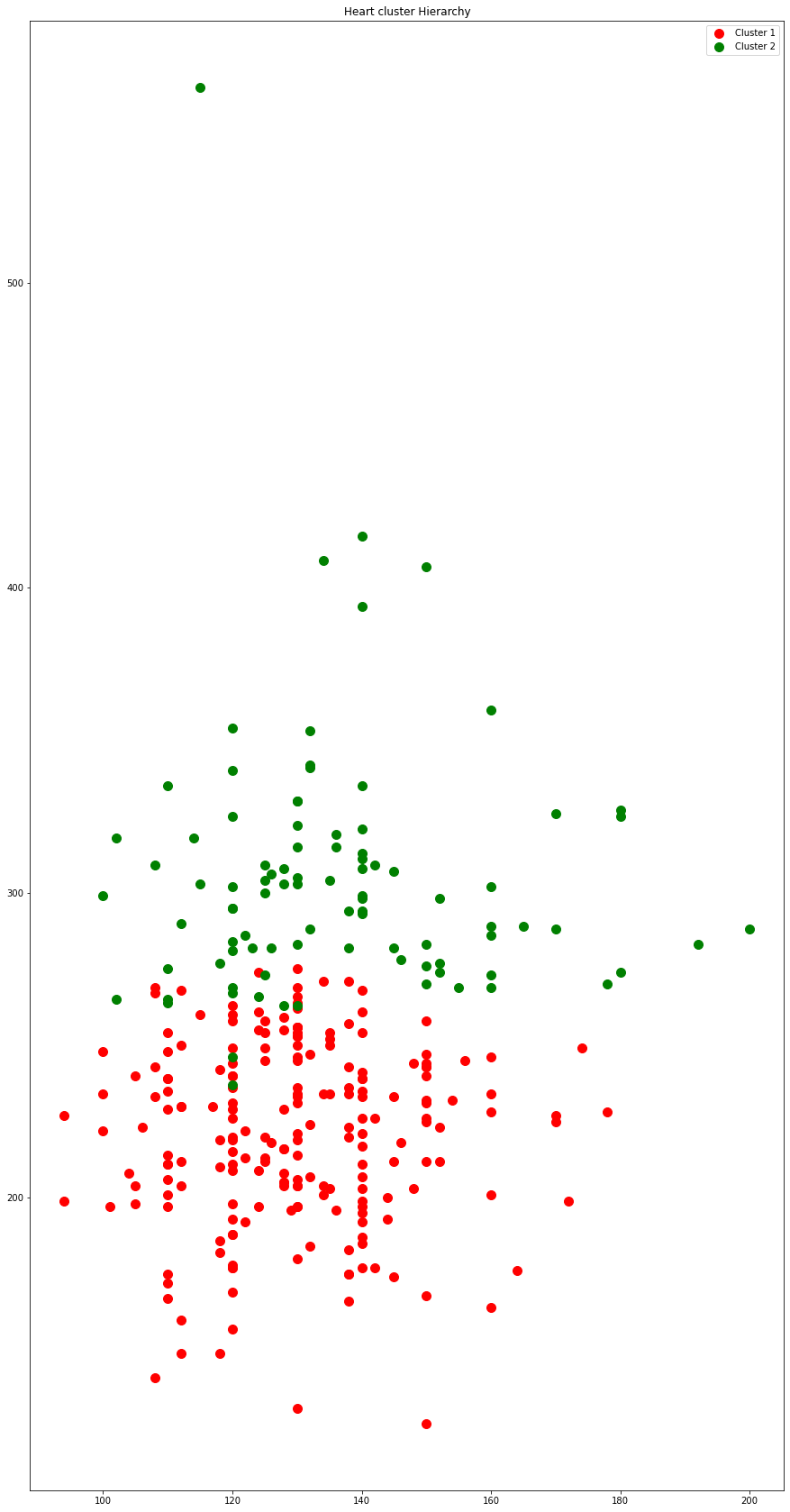
****

**Hieratical clustering**

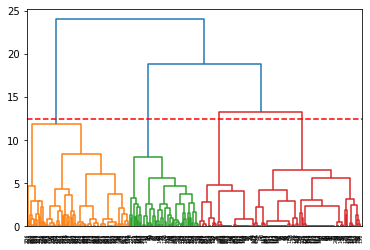
At the Hieratical clustering can be shown as dendrogram. We start at the bottom with some certain number of points each one of them assigned to different cluster. The closest ones are merged till at the end we have only one cluster at the top.

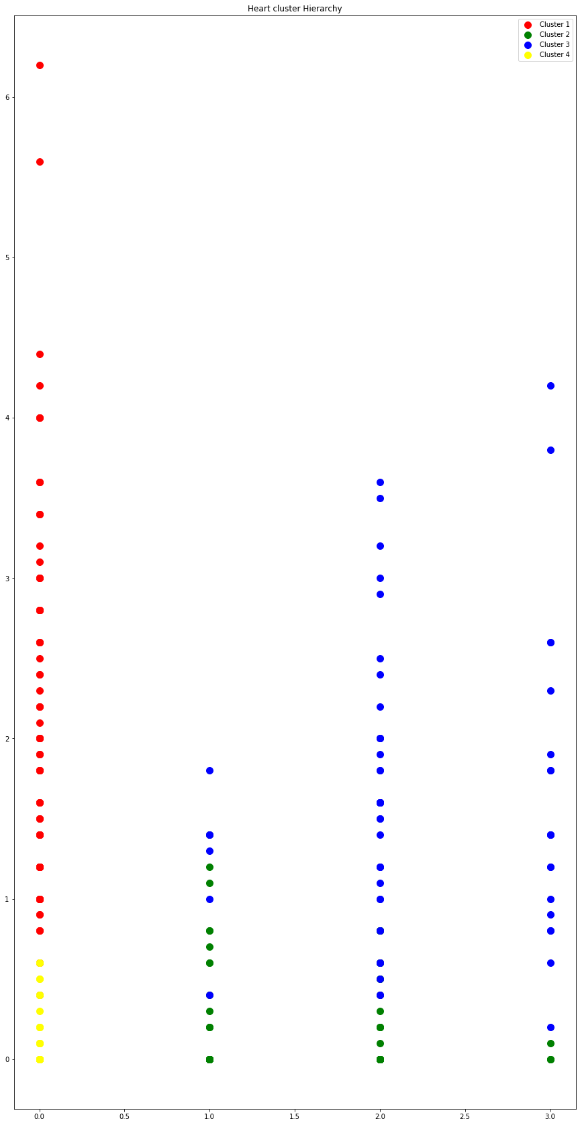
**First code is was by taking row 3,4 and 7 and by getting the dendrogram the optimal number of clusters was 2 so the visualizing is:**



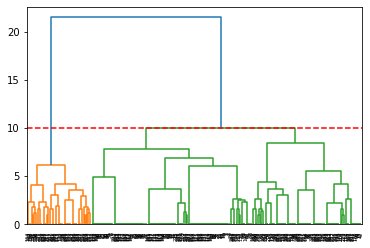


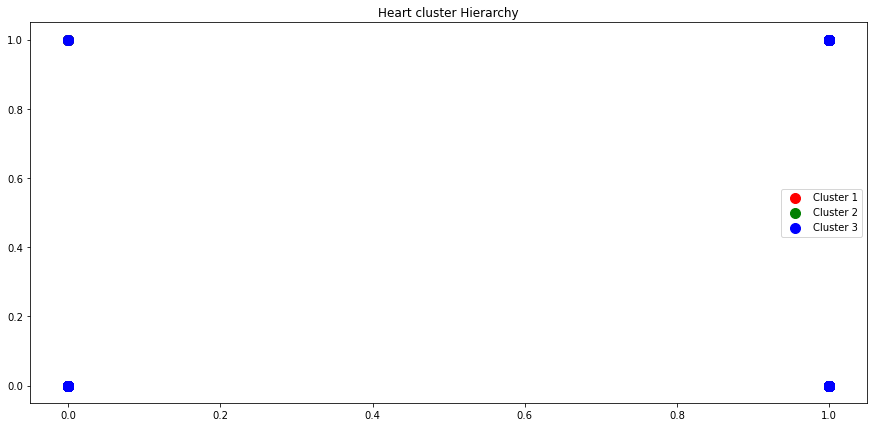
**second code is was by taking row 2,9,10 and 12 and by getting the dendrogram the optimal number of clusters was 4 so the visualizing is:**

****

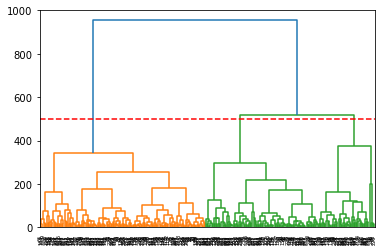
****

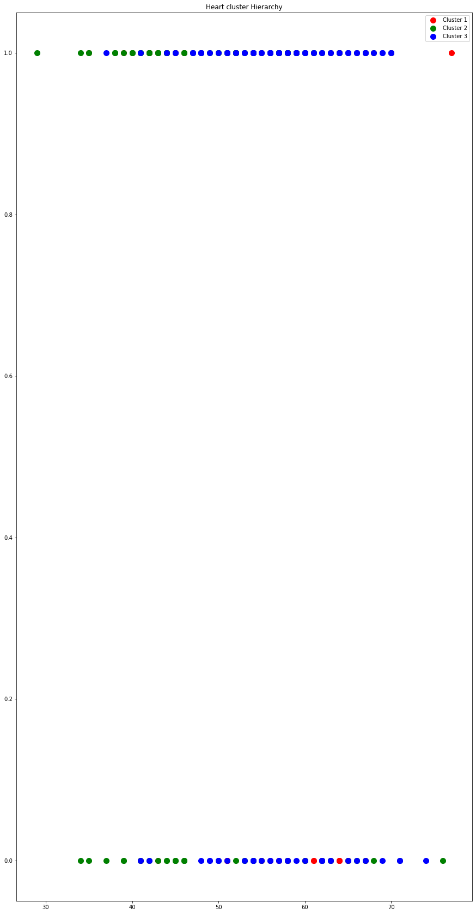
**third code is was by taking row 1,5,6,8 and 11 and by getting the dendrogram the optimal number of clusters was 3 so the visualizing is:**

****

****

**The last code is was by taking row all the columns and by getting the dendrogram the optimal number of clusters was 3 so the visualizing is:**

****

****