GLA University, Mathura-281406 (U.P.) INDIA
Research Paper 14 December 2019

Clustering Heart Disease Patient Data

Aman Pratap Singh[1], Arpit Agnihotri[2], Ishan Agrawal[3], Vatsal Raj Krishna[4]

[1] Department of Computing Engineering and Applications, GLA University, Mathura, India
[2] Department of Computing Engineering and Applications, GLA University, Mathura, India
[3] Department of Computing Engineering and Applications, GLA University, Mathura, India
[4] Department of Computing Engineering and Applications, GLA University, Mathura, India

***Abstract*—** Every year about nineteen million individuals die from heart diseases worldwide. A heart patient shows many symptoms and it's very tough to attribute them the heart disease after so many steps of disease progression. Data mining is a solution to extract a hidden pattern from the clinical dataset, which are applied to a info during analysis. All offered algorithms in clustering technique are compared to every alternative to attain the very best accuracy. To additional increase the correctness of the answer, the dataset is preprocessed by totally different clustering algorithms. The two necessary tasks that are required for the event of clustering are K-means clustering and Hierarchical clustering. In K-means clustering the initial point selection effects on the results of the algorithm, each within the variety of clusters found and their centroids too. Ways to reinforce the K-means cluster formula is mentioned. With the assistance of those ways efficiency, accuracy and performance are improved. So, to boost the performance of clusters the normalization could be a pre-processing stage i.e. employed to reinforce the Euclidean distance by calculating a most nearer centers, which help in reducing number of iterations which is able to reduce the machine time as compared to k-means clustering algorithm. Finally, again the clusters are made with hierarchical clustering where clusters are formed either in bottom-up approach or top-down approach. The hierarchical clustering technique adopted performs comparatively well for the sake of doctors as it gives multiple patients in each group so they can compare treatments.

***Keywords*—** Data mining, K-means clustering, Hierarchical Clustering

I. INTRODUCTION

Among all the harmful diseases, heart attacks are one of the most universal diseases. Medical practitioners conduct a large amount of surveys on heart diseases and collect information of heart patients, their disease progression and symptoms. So there are various things that can lead a heart attack. Thus, there's valuable info hidden in their dataset to be taken out. Data processing is that the technique of retrieving hidden info from an oversized set of information. It helps researchers to gain each and every profound insights of new understanding and novel of huge medical datasets. The foremost necessary or say the most important goals of knowledge mining are prediction and outline of diseases. It's earned through the process of a group of variables (attributes) within the dataset and discovering the long run states of remainder variables.

Extracting key information from a large amount of data is simply called data mining. The more appropriately this data mining is nothing but knowledge mining. Knowledge mining doesn't have the exact meaning of data mining, as it doesn't reflect the emphasis on the extraction of data from a large amount of data. In recent days, this data mining, that is "data" and "mining" has become very much popular among researchers. To carry out this data mining process, the following sequence of steps is very much important.

1. Data Web
2. Information Retrieval(Resource discovery)
3. Information Extraction(Selection/Preprocessor)
4. Generalization(Pattern Recognition)
5. Analyze(Validation)
6. Knowledge

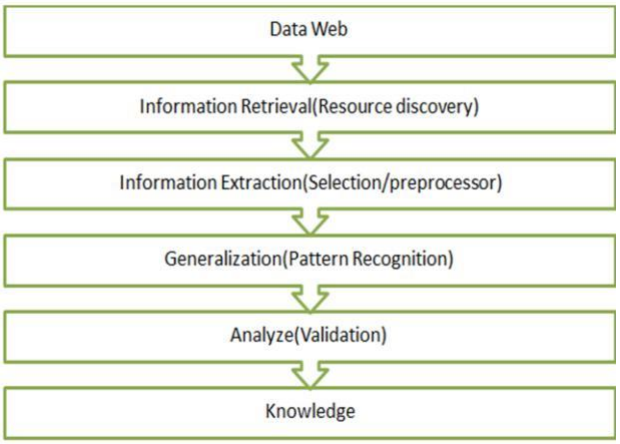


Figure 1: Data mining process

A. Clustering

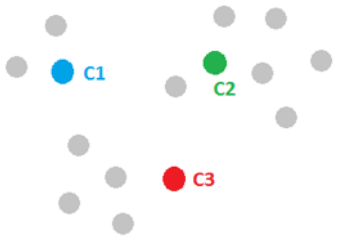
Clustering analysis is a method used widely in the data mining community. It is a data mining technique used to place the data elements into their related groups called as clusters. This technique helps in summarizing a very large data set X with much small set C= {ci| i=1, 2, 3, 4, 5, 6 …….k} of the representative points called as centroids and a membership map γ: X→C relating each point of X to its representative in C. Various clustering algorithm like hierarchical, EM algorithm, k-means etc. are used to make a set C of Representatives. In this paper, K-means clustering and Hierarchical clustering are used.

B. K-Means Clustering

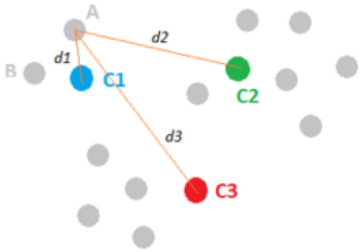
The K-means clustering is a simple as well as popular unsupervised machine learning algorithms. An unsupervised algorithms make inferences from datasets using input vectors without referring to known or labeled, outcomes. The K-means algorithm works by identifying the k number of centroids and then allocating every data point to the nearest cluster, while keeping the centroids as small as possible.

Let us understand the working of K-means clustering step-wise:-

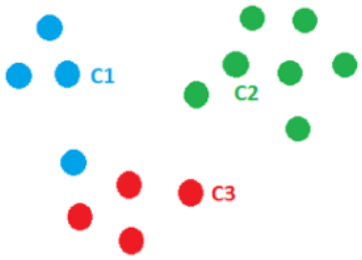
Step 1: *First Initialize cluster centers*
Here we randomly pick three points, say C1, C2 and C3, and label them with blue, green and red color separately to represent the cluster centroids.



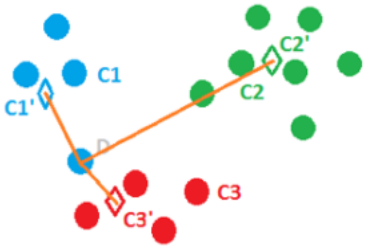
Step 2: *Assign observations to the closest cluster center*



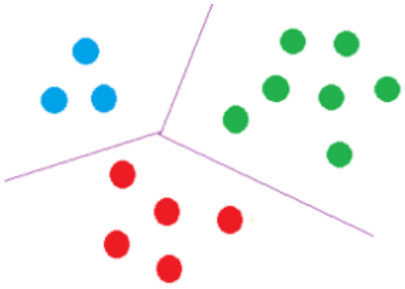
After getting these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center by calculating the Euclidean distance. This process will assign all the points to the corresponding clusters and leads to the following figure



Step 3: *Revise the cluster centroids as mean of assigned observations*
Now, we need to update the cluster centers For example, we can find the center mass of the blue cluster i.e. C1 by adding over all the blue points and dividing by the total number of points, which is four here. And the resulted center mass C1', represented by a blue diamond, is our new centroid for the blue cluster. Similarly, we can find the new centroids of C2' and C3' for the green and red clusters.



Step 4: *Repeat Step 2 and Step 3 until convergence*
The last step of k-means algorithm is just to repeat the step 2 and step 3. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centroids, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centroids, and updating the cluster centroids until convergence. Finally, we may get a solution like the following figure. Well done!



C. Hierarchical Clustering

The hierarchical clustering Technique is one of the popular Clustering techniques in Machine Learning. Hierarchical clustering starts by treating each observation as a separate cluster of dataset.

A dendrogram is a diagram that shows the hierarchical relationship between different objects of a dataset. It is most commonly created as an output of hierarchical clustering to display step-wise process. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

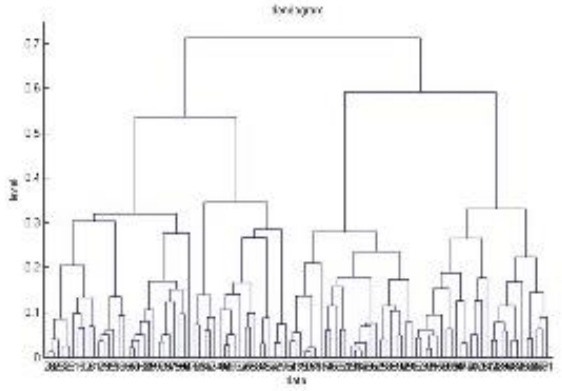


Figure 2: A Dendrogram

D. Heart Disease

The heart is a very important organ or part of our body. Life is itself dependent on proper working of heart. The operation is heart is to hump blood to all the parts of body with help of arteries and also clean the blood carried to it by veins. If operation of heart is not appropriate, it will affect the other body parts of human such as kidney, brain, etc. Heart is nothing more than a pump, which pumps blood through the body. If the circulation of blood in the body is inefficient the body parts like brain suffer and if the heart stops working altogether, death occurs within minutes. Life is utterly reliant on the proficient operation of the heart. The term Heart disease refers to the illness of the heart & blood vessel system in it. There are a number of factors that amplifies the risk of Heart disease such as the Family history of heart disease, Poor diet, Cholesterol, Smoking, High blood pressure, high blood cholesterol, Hypertension, Physical inactivity, Obesity, etc.

E. Symptoms of a Heart Attack

Symptoms of a heart attack are Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone. Anxiety burning at back, jaw, throat, indigestion, or arm Fullness or choking feeling (may feel like heartburn). Some of the common indications are Sweating, nausea, vomiting or dizziness that also includes anxiety, extreme weakness, or shortness of breath, rapid or irregular heartbeats.

Rest of the paper is organized as follows, section I contains the introduction of the paper, section II contains the Background and Literature survey of the Paper, section III contains the Proposed research methodology of the paper in this section we tried to explain our proposed algorithm.

II. BACKGROUND AND LITERATURE SURVEY

In any nation to have a progressive development in all the sectors, it is very important and particularly to have attention towards the health of their population. According to that, heart diseases and myocardial ischemic (Joe-Air Jiang et al. 2006) are the most common heart diseases which can lead to serious conditions and cause of death in most of the industrialized countries (Minami et al. 1999). These forms of heart diseases can be detected from the information taken from the so many heart attributes and they normally give the data related with health conditions of the patients. It is very much essential to enhance the patients living condition and treatment.

In this paper, authors proposed k-means clustering algorithm and hierarchical clustering to distinguish patients data in different clusters on the basis of age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rates achieved, resting electrocardiographic result, fasting, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise, number of major vessels colored by fluoroscopy and Thalassemia. Patients with similar characteristics might respond to the same treatments, and doctors would benefit from learning about the outcomes of patients similar to those they are treating.

Therefore, a comprehensive comparison of clustering algorithms practically provides an insight into performances. After the k-means algorithm we also performed hierarchical clustering algorithm. This comparison plays great importance to medical practitioners who desire to predict heart failure at a proper step of its progression. In any nation to have a progressive development in all the sectors, it is very important and particularly to have attention towards the health of their population. According to that, heart diseases and myocardial ischemic (Joe-Air Jiang et al. 2006) are the most common heart diseases which can lead to serious conditions and cause of death in most of the industrialized countries (Minami et al. 1999).

These kinds of heart diseases can be detected from the information taken from the so many heart attributes and they normally give the data related with health conditions of the patients. It is very much essential to enhance the patients living condition and treatment.

III. PROPOSED RESEARCH METHODOLOGY

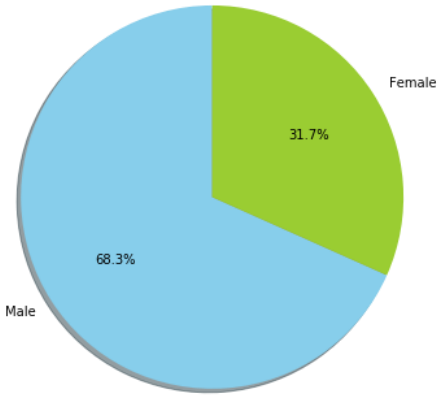
In k-mean clustering algorithm, the goal is to find groups of data (data is unlabeled) and after that functions are clustered based on feature similarity by using Euclidian distance formula. In this paper, the quality of clusters is increased by enhancing the Euclidian distance formula. The enhancement

that has to do will be based on normalization. Normalization which is a pre-processing technique will enhance the accuracy and efficiency of clusters by calculating best distances from the dataset which will result in more accurate center points and as a result best cluster results are formed, the feature which is added is for calculating normal distance metrics on the basis of normalization.

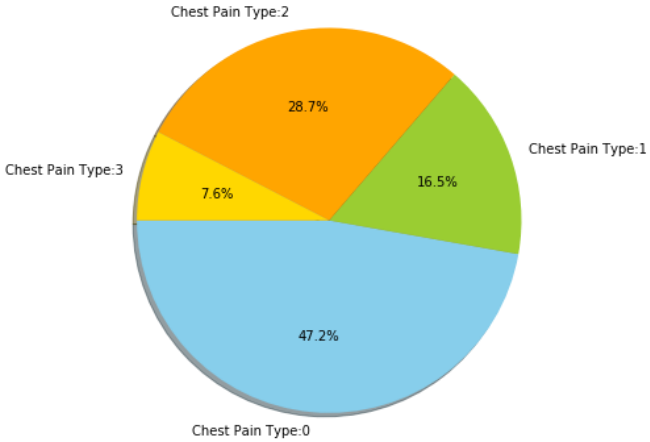
IV. EXPERIMENTAL RESULTS

In this paper Heart disease, dataset is used for the research process and prediction analysis. Before applying k-means clustering and hierarchical clustering algorithms let us study about the variables of datasets to have better understanding of results, that we will be getting in last of the project.

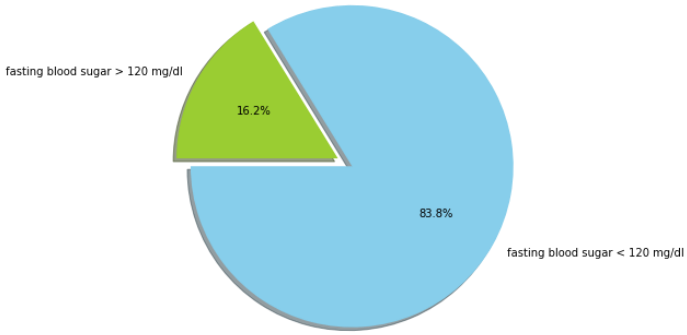
1. Sex



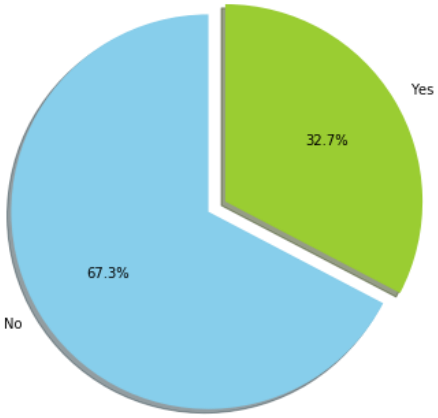
2. Chest Pain Type



3. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)



4. exang: exercise induced angina (1 = yes; 0 = no)



We have also checked for the null values in every column of a dataset. Missing data can reduce the statistical power of a result and that can lead to produce biased estimates, leading to invalid conclusions. But the dataset we have doesn't have any null values.

if (any(is.na(heart_disease["column_name"]))) {next}

After checking for all the columns we moved further in our project and checked if all the values are numeric values so that we can apply clustering algorithms and have better results.
If we do not have numeric values than we have to apply label encoding. Label Encoding ensures to give a numeric value to a string value of same types.

lapply (heart_disease, class)

It is important to conduct some exploratory data analysis to familiarize ourselves with the data before clustering. This will help us learn more about the variables and make an informed decision about whether we should scale the data. Because k-means and hierarchical clustering measures similarity between points using a distance formula, it can place extra emphasis on certain variables that have a larger scale and thus larger differences between points. Exploratory data analysis helps us to understand the characteristics of the patients in the data. We need to get an idea of the value ranges of the variables and their distributions. This will also be helpful when we evaluate the clusters of patients from the algorithms. Are there more patients of one gender? What might an outlier look like? But with the help of this above information and different checking we can proceed further.

Step 1: Applying first round of k-means clustering algorithm to the dataset.

A scatter heart disease dataset is first loaded in an R Studio. Once we've figured out if we need to modify the data and made any necessary changes, we can now start the clustering process. For the k-means algorithm, it is necessary to select the number of clusters in advance. It is also important to make sure that your results are reproducible when conducting a statistical analysis. This means that when someone runs your code on the same data, they will get the same results as you reported. Therefore, if you're conducting an analysis that has a random aspect, it is necessary to set a seed to ensure reproducibility. Reproducibility is especially important since doctors will potentially be using our results to treat patients. It is vital that another analyst can see where the groups come from and be able to verify the results.

So, we have to select the value of 'k' to make 'k' number of centroids and then allocating every data point to the nearest cluster. In this round we are taking k=5, i.e. we are making 5 clusters centroid in this step.

Step 2: Applying second round of k-means clustering algorithm to the dataset.

Because the k-means algorithm initially selects the cluster centers by randomly selecting points, different iterations of the algorithm can result in different clusters being created. If the algorithm is truly grouping together similar observations (as opposed to clustering noise), then cluster assignments will be somewhat robust between different iterations of the algorithm.

With regards to the heart disease data, this would mean that the same patients would be grouped together even when the algorithm is initialized at different random points. If patients are not in similar clusters with various algorithm runs, then the clustering method aren't picking up on meaningful relationships between patients.

We're going to explore how the patients are grouped together with another iteration of the k-means algorithm. We will then be able to compare the resulting groups of patients.

Step 3: Comparing patient k-means clusters results

It is important that the clusters resulting from the k-means algorithm are stable. Even though the algorithm begins by randomly initializing the cluster centers, if the k-means algorithm is the right choice for the data, then different initializations of the algorithm will result in similar clusters.

The clusters from different iterations may not be exactly the same, but the clusters should be roughly the same size and have similar distributions of variables. If there is a lot of change in clusters between different iterations of the algorithm, then k-means clustering is not a good choice for the data.

It is not possible to validate that the clusters obtained from an algorithm are ground truth are accurate since there is no true labeling for patients. Thus, it is necessary to examine how the clusters change between different iterations of the algorithm. We're going to use some visualizations to get an idea of the cluster stabilities. That way we can see how certain patient characteristics may have been used to group patients together.

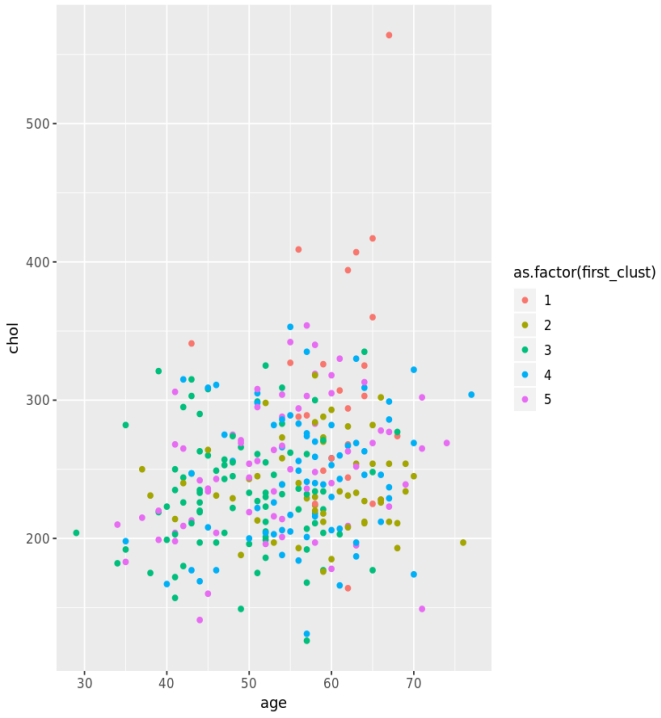


Figure 3: K-Means Clustering Round 1 Result

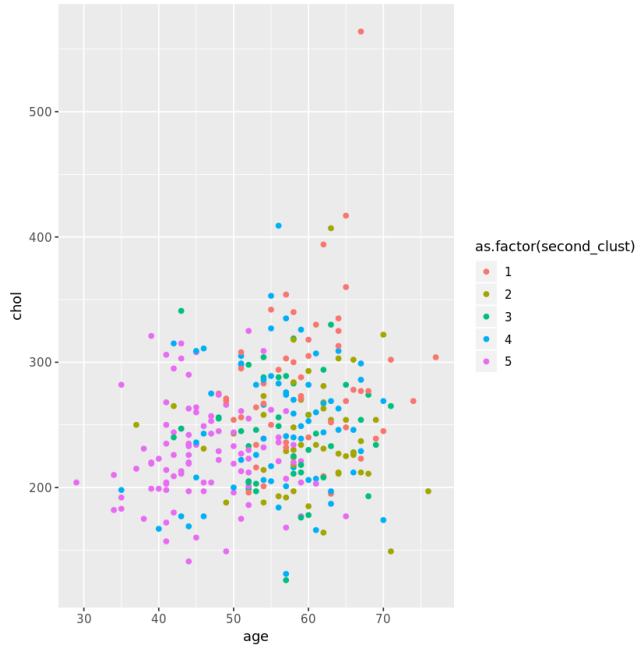


Figure 4: K-Means Clustering Round 2 Result

Step 4: Applying first round of Hierarchical clustering algorithm to the dataset.

An alternative to k-means clustering is hierarchical clustering. This method works well when the data has a nested structure. It is possible that the data from heart disease patients follows this type of structure. For example, if men are more likely to exhibit certain characteristics, those characteristics might be nested inside the gender variable. Hierarchical clustering also does not require the number of clusters to be selected prior to running the algorithm.

Clusters can be selected by using the dendrogram. The dendrogram allows one to see how similar observations are to one another and are useful in selecting the number of clusters to group the data. It is now time for us to see how the hierarchical clustering algorithm groups the data of the dataset. In this step we are applying single linkage hierarchical clustering algorithm to make clusters of dataset.

Step 5: Applying second round of Hierarchical clustering algorithm to the dataset.

In hierarchical clustering, there are multiple ways to measure the dissimilarity between clusters of observations. Complete linkage records the largest dissimilarity between any two points in the two clusters being compared. On the other hand, single linkage is the smallest dissimilarity between any two points in the clusters. Different linkages will result in different clusters being formed.

We want to explore different algorithms to group our heart disease patients. The best way to measure dissimilarity between patients could be to look at the smallest difference between patients and minimize that difference when grouping together clusters. It is always a good idea to explore different dissimilarity measures. Let's implement hierarchical clustering using a new linkage function.

Step 6: Comparing patient Hierarchical clusters results

It is important that the clusters resulting from hierarchical algorithm are stable. Even though the algorithm begins by randomly initializing the cluster centers, if the hierarchical algorithm is the right choice for the data, then different initializations of the algorithm will result in similar clusters.

The clusters from different iterations may not be exactly the same, but the clusters should be roughly the same size and have similar distributions of variables. If there is a lot of change in clusters between different iterations of the algorithm, then hierarchical clustering is not a good choice for the data.

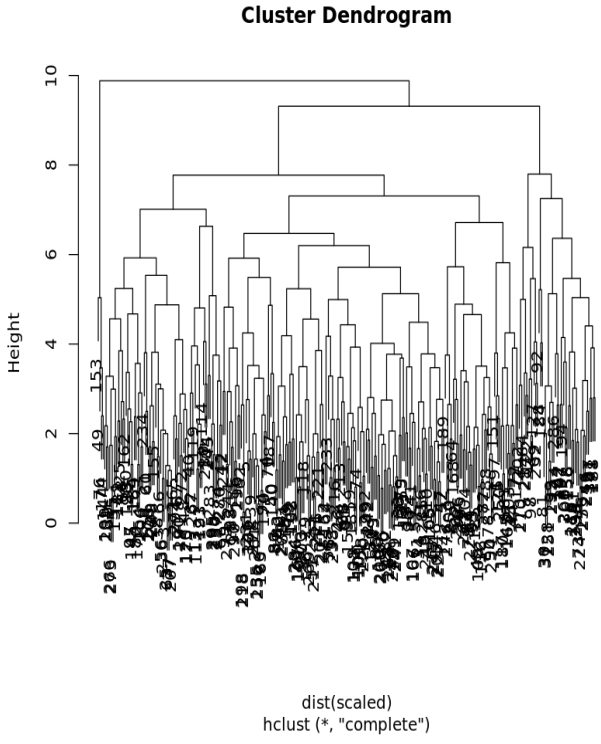


Figure 5: Hierarchical Clustering Round 1 Result

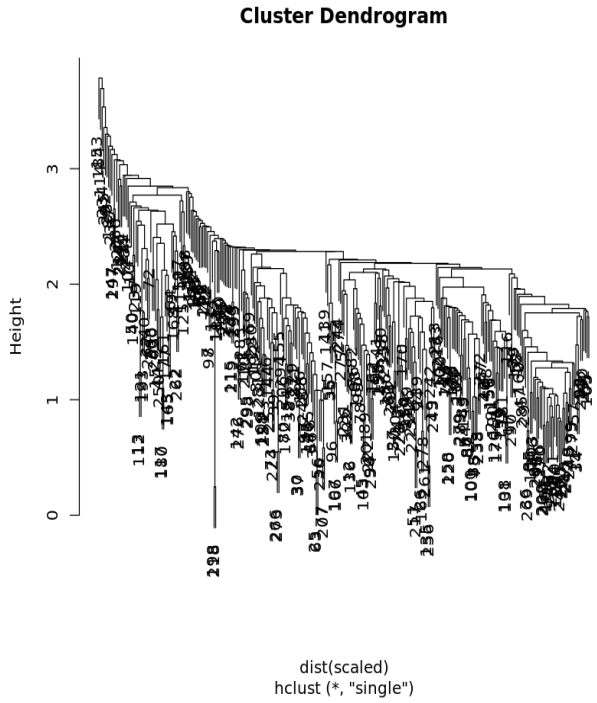


Figure 6: Hierarchical Clustering Round 2 Result

Figure 7: Plotting Age and Cholesterol

Step 7: Comparing Clustering results

The doctors are interested in grouping similar patients together in order to determine appropriate treatments. Therefore, they want to have clusters with more than a few patients to see different treatment options. While it is possible for a patient to be in a cluster by themselves, this means that the treatment they received might not be recommended for someone else in the group.

As with the k-means algorithm, the way to evaluate the clusters is to investigate which patients are being grouped together. Are there patterns evident in the cluster assignments or do they seem to be groups of noise? We're going to examine the clusters resulting from the two hierarchical algorithms.

Step 8: Visualizing the Cluster contents

In addition to looking at the distributions of variables in each of the hierarchical clustering run, we will make visualizations to evaluate the algorithms. Even though the data has more than two dimensions, we can get an idea of how the data clusters by looking at a scatterplot of two variables. We want to look for patterns that appear in the data and see what patients get clustered together. So here are the 2 plots which will help us to visualize the clustering results.

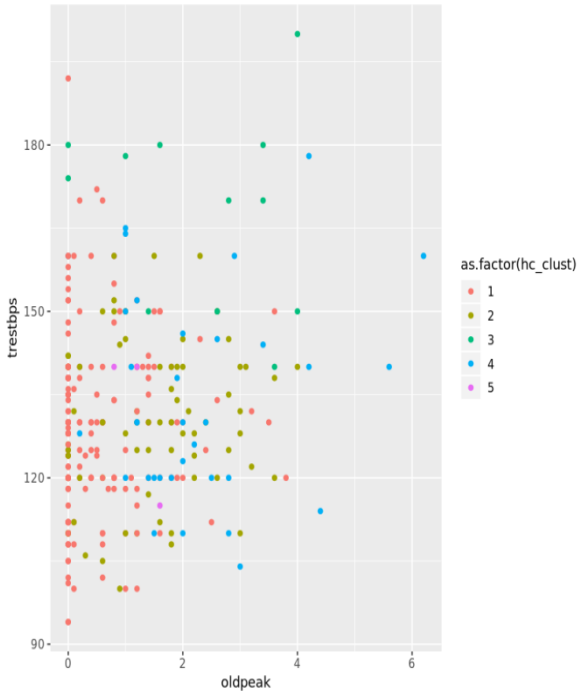


Figure 8: Plotting Oldpeak and Trestbps

V. CONCLUSION

Now that we've tried out multiple clustering algorithms, it is necessary to determine if we think any of them will work for clustering our patients. For the k-means algorithm, it is imperative that similar clusters are produced for each iteration of the algorithm. We want to make sure that the algorithm is clustering signal as opposed to noise.

For the sake of the doctors, we also want to have multiple patients in each group so they can compare treatments. We only did some preliminary work to explore the performance of the algorithms. It is necessary to create more visualization and explore how the algorithms group other variables. Based on the above analysis, are there any algorithms that you would want to investigate further to group patients? Remember that it is important the k-mean algorithm seems stable when running multiple iterations.

References

[1] https://www.kaggle.com/ahmadjaved097/classifying-heart-disease-patients

[2] https://healthcare.ai/step-step-k-means-clustering/

[3]de Carvalho Junior, Helton Hugo, et al. "A heart disease recognition embedded system with fuzzy cluster algorithm." Computer methods and programs in biomedicine 110.3 (2013): 447-454.

[4]K.Rajalakshmi,Dr.S.S.Dhenakaran,N.Roobin"Comparative Analysis of K-Means Algorithm in Disease Prediction". International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 7, July 2015.

[5]Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma," Survey on Different enhanced K-means Clustering Algorithm", International Journal Of Engineering Trends And Technology, Vol. 27 ,No. 4-September 2015.

[6] R. Alizadehsani, J. Habibi, M. Hosseini, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.Alizadehsani, "A Data Mining Approach for Diagnosis of Coronary Artery Disease," Elsevier, 2013.

Authors Profile

Aman Pratap Singh, pursuing B.Tech in computer Engineering and Applications (Third Year) in GLA University.

Arpit Agnihotri, pursuing B.Tech in computer Engineering and Applications (Third Year) in GLA University.

Ishan Agrawal, pursuing B.Tech in computer Engineering and Applications (Third Year) in GLA University.

Vatsal Raj Krishna, pursuing B.Tech in computer Engineering and Applications (Third Year) in GLA University.