INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1730)
Available online at: https://www.ijariit.com

Heart disease clustering using K-Mean analysis

Gomanth D Reddy
dgreddi2000@gmail.com
SRM University, Chennai, Tamil Nadu

ABSTRACT

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering usually used to classify data into structures that are more easily understood and manipulated.This research usesK-Means Clustering method and is imposed on Heart Disease Dataset. In this paper, the risk factors that causes heart disease is considered and predicted using K-means and the analysis is done in Jupiter notebook using an existing data available online for heart disease. The dataset consists of 209 people along with 8 attributes named as age, blood_pressure, chest_pain_type ECG_in_rest, blood glucose level, heart_rate and four types of chest pain. The prediction of the heart disease is done by using K-means clustering algorithm along with data analytics and visualisation tools. This paper shows the clustering work done on the taken data. The clustering results is show in the notebook and are accurate.

Keywords–*Clustering, K-Means Clustering*

1. INTRODUCTION

Clustering is a method of grouping data on a database based on the given attributes. All the clustering result are shown to the enduser for him/her to know how the method works on the database. Particular data class is not required to be grouped to perform Clustering. Even clustering can be used to label the unknown data class. Therefore, clustering is often classified as an unsupervised learning method .One such method of data grouping is K-Means Clustering. K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clusteredbased on feature similarity.The prediction of any heart disease is a very arduous and challenging aspect in the medical industry, it can be based on many factors which include the patients history, physical tests and the symptoms shown by the patient .There are many factors that influenceheart diseases such as levels of cholesterol in the body, work place, food habits, obesity, historyof diseases in the family, BP, and their living environment.In predicting heart disease manyMachine learning algorithms have a big role to play . The new methods and algorithms makes it easy to work with unstructured data which is growing daily. This paper shows us the methodsused to predict the heart disease which have been worked in the Jupyter Notebook.

2. HEART DISEASES

This is the most widespread disease across the world and is a deadly disease. This is caused duethe decrease of blood flow into various organs of the human body. There are many symptoms to identify if a person has a heart disease such as swollen legs , weakness in the body and uneven breathing ,etc. Identification of the particular heart disease is important to treat the patient as soon as possible before its becomes more complicated and life taking. In the present scenario theman-power available for the treatment and diagnosis process is limited so this increases the risk of people dying of heart diseases. The techniques used in this paper identifies or predicts the heart disease of the person based on the attribute values they fill in , so this helps us reduce the risk of it affecting the human body.

3. DATA CLUSTERING

This is a part of data mining where we identify the similarities between the existing data and cluster them according to the necessary requirements. This is done using many mathematical andstatistical techniques. The data is clustered based on certain criteria and the results are given to the end user to get an idea of the functioning of the database. No particular data class is required to perform clustering on the data which makes it easy to code. The clustering process can be used to label the data classes which are unknown. There are various clustering methods which include unsupervised learning and descriptive methods such as matlabplot etc. The objects are grouped based on high similarity between them and they differ highly from the other clusters. A good clustering method can yield great clusters with similar objects .

4. K-MEANS CLUSTERING

This is a non-hierarchical method and partitions the dataset into several clusters k. This divides the data into different groups and accepts new data without their respective class labels. In this type of clustering the data with same characteristics are grouped into one cluster and the rest ofthe data is grouped into another cluster groups. The variating of data in the clusters are minimal which makes the K-Means algorithm useful and relatively fast and is widely used in the industry.

The K-Mean Clustering is done in the following way:

- The number of clusters is chosen first
- K random points are selected from the data as centroids
- points are assigned to the closest cluster centroids
- The newly formed clusters centroids are recomputed
- The steps 3 and 4 must be repeated.

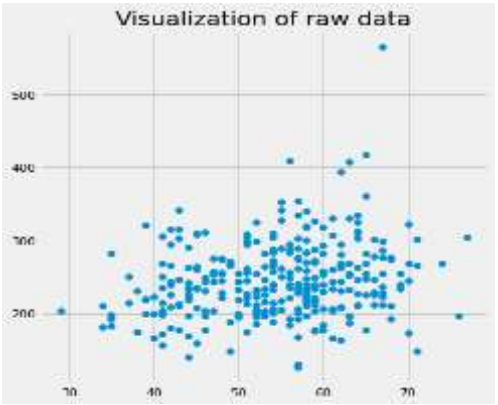
Dataset

Here we have the dataset which is available online named(Cleveland Heart Disease Dataset). This consists of a total of 76 features and 303 patients data which is collected by the organisation. First we read the data after which we identify the incomplete data using get_dummies and remove the data . We now have 209 samples to train our model . These 209 samples have some attributes such as age, chol, fbs, thalach, etc. In the dataset age is the most important factor as diseases as prone to develop at the adolescence stage. This dataset has onlymale patients as we are focusing on only coronary diseases which are at a higher risk to occur in male patients then in female patients.

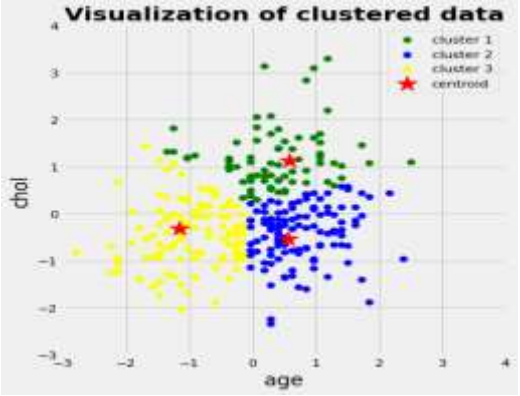
age	trestbps	chol	fbs	thalach	exang
0	63	145	233	1	150
1	37	130	250	0	187
2	41	130	204	0	172
3	56	120	236	0	178
4	57	120	354	0	163

5. RESULTS AND DISCUSSION

The model we have created using Jupyter notebook will help us predict the heart disease of theparticular patient . We have analysed the data using K-Means Clustering and found out hidden patterns in the preexisting dataset to help us speed up the process of identification of the heartdisease . We have clustered the data and found out that there are three types of heart diseases namely typical anigna pain , non-anginal pain , asymptomatic pain . The raw data is visualised using scatter plot.



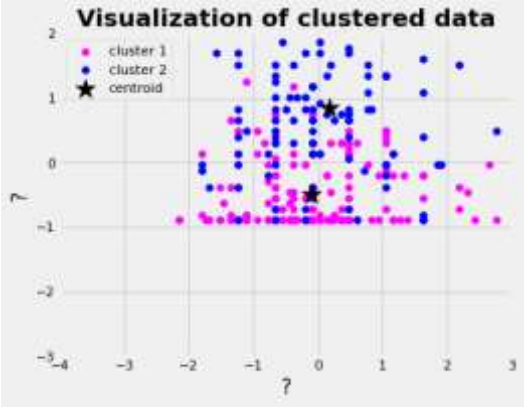
We then cluster the data by giving them labels and fig size: chol/age:



Then we describe the clustered data:

	age	chol	cluster
count	303.000000	303.000000	303.000000
mean	54.366337	246.264026	1.115512
std	9.082101	51.830751	0.811664
min	29.000000	126.000000	0.000000
25%	47.500000	211.000000	0.000000
50%	55.000000	240.000000	1.000000
75%	61.000000	274.500000	2.000000
max	77.000000	564.000000	2.000000

Now we use K-Means to cluster the data :



The clusterssurround thecentroids .

6. FUTURE ADVANCES AND CONCLUSION

From the analysis of the data using the K-Means Algorithm we get to know that age is the mainfactor in the prediction of the disease as the centroids of the plots where we consider age as an attribute has the biggest concentration and similarities. The model predicts the disease using themajor factors which the four types of chest pain. The K-Means clustering helps us predict the happening of a chest pain in advance and is thus a very simple and popular clustering algorithmused in the industry .

7. REFERENCES

[1] V. Manikantan & S.Latha,"Predicting the Analysis of Heart Disease Symptoms Using MedicinalData Mining Methods", International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.

[2] Dr.A.V.Senthil Kumar, "Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering", International Journal of Advanced Trends in Computer Science andEngineering, Volume-4, No.6, pp.07-18, 2015.

[3] Uma.K, M.Hanumathappa, "Heart Disease Prediction Using Classification Techniques withFeature Selection Method", Adarsh Journal of Information Technology, Volume-5, Issue-2, pp.22-29, 2016

[4] Himanshu Sharma, M.A.Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms:A Survey",International Journal on Recent and Innovation Trends in Computing andCommunication,Volume5,Issue-8,pp.99-104, 2017.

[5] S.Suguna, Sakthi Sakunthala.N ,S.Sanjana, S.S.Sanjhana, "A Survey on Prediction of HeartDisease using Big data Algorithms", International Journal of Advanced Research in Computer Engineering & Technology,Volume-6,Issue-3,pp.371-378,2017.

[6] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure,"Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, 2011.

[7] J.Mourão-Miranda,A.L.W.Bokde,C.Born,H.Hampel,and M. Stetter, "Classifying brain states and determining the discriminatingactivationpatterns:supportvectormachineon functionalMRIdata,"NeuroImage,vol.28,no.4,pp.980–995, 2005.

[8] S.Ghwanmeh,A.Mohammad,andA.Al-Ibrahim,"Innovative artificial neural networks-based decision support system for heartdiseasesdiagnosis,"JournalofIntelligentLearningSystems andApplications, vol. 5, no. 3, pp. 176–183, 2013.

[9] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal ofComputer Science Issues, vol. 8, no. 2, pp. 150– 154, 2011.

[10]K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," International Journal of Computer Applications, vol. 19, no. 6, pp. 6–12, 2011.

