JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

An Expert System For Heart Disease Prediction Applying On Different Clustering Approach

[1]Dr.V.Poornima , [2]Dr.M.Rahimal Beevi and S.Meenakshi

[1]Assistant Professor, Department of Computer Science, SRMIST, Ramapuram,Chennai,poornimasudhaagar@yahoo.com
[2]HOD,Department of Computer Science, Mohamed Sathak College of Arts and Science,Chennai,rahimaarifeen@gmail.com
[3]Assistant Professor, Department of Computer Science, New Prince Shri Bhavani Arts and Science College,Chennai,meenasmail1979@gmail.com

ABSTRACT
Heart attacks are believed to be the most common of all hazardous disorders. Medical professionals perform several surveys on Heart Disease(HD) and gather data on patients, their illness progression, and symptoms.In many levels of disease progression, it is difficult to connect a heart patient's symptoms to the heart disease. In this study, Data Mining (DM) is applied to a database in order to find a hidden trend in a clinical dataset. Clustering is a habitual DM technique that organizing data elements based on their commensurable criterion. This study revealed how to capture clusters and establish the new centroid utilizing K-Means (KM) , Canopy Clustering (CC) algorithms and Farthest First Algorithms. The KM technique is a broadly used clustering technique that is employed in a extensive range of scientific and industrial applications. Canopy Clustering is a straight forward and rapid approach for accurately grouping items into clusters. Farthest First Techniques (FFT) is fitting for the massive dataset and its fabricates for the non-uniform cluster. In the end, the execution of three algorithms were investigated in this work. The American Heart Association (AHA) provided a dataset of 50 participants to examine the heart disease dataset.

Keywords: Clustering Techniques, Heart Disease Prediction, Data Mining techniques, AHA Dataset, Weka Tool

I. INTRODUCTION

Nowadays, HD is the foremost reason for the death worldwide. HD affects a large number of people each year brutally all over the map consistent with the WHO, 17.5 million people died from HD globally in 2012 [1]. Human mortality can be reduced if cardiac disease can be predicted. In the subject of health, information technology is critical [2]. The practise of extracting hidden information from a vast group of databases is known as data mining.It assists researchers in gaining both profound insights and unique ideas from vast medical datasets [3]. Selecting, analysing, preparing, applying, interpreting, and assessing the results are all part of the data mining process [4]. Clustering is a DM approach that clusters the data elements based on their resemblance[5]. This research work aims to give a comprehensive analysis of the K Means algorithm , Canopy clustering (CC)and Farthest Firs methods for predicting heart disease. These two algorithms use the American Heart Association(AHA) dataset, and their clusters are estimated. The weka tool was used to investigate the results of various algorithms.

II. RELATED WORK

Clustering is an important activity in data investigation that seeks to uncover data structures with inherent state by assembling data items into comparable groups and representing data in classes; as a result, it is referred to as unsupervised classification or observational learning [6].The primary objective of clustering process is to assemble similar and dissimilar objects into the identical clusters and then separate them. Objects in one cluster are the same as those in another, but they are not the same as those in other clusters[7]. Clustering analysis' fundamental purpose is to arrange comparable and dissimilar

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

objects into the same clusters and separate clusters, accordingly. Objects in one cluster are the same as those in another, but they are not the same as those in other clusters.. Clustering approaches [8] do not know the semantics of the classes beforehand.

KM is a cluster analysis technique that divides the set of items into K groups based on their properties. Using the Euclidean distance formula and the related cluster centroid, the sum of squares of distances between data is minimized. The findings of the study demonstrate that combining clustering produces promising outcomes with the highest accuracy rate and resilience [9]. There are three ways that a huge data set can be created: (1) There might be a lot of elements in the data set, (2) each element could contain a lot of features, and (3) there could be a lot of clusters to find. When the problem is enormous in all three of these ways at the same time, clustering is an effective strategy. The main idea is to cluster data in two stages: a rough and quick step that sunder the data into "canopies," and a more thorough final stage that only conducts immoderate distance measurements between points that occur under the same canopy. [10].

III. CLUSTERING ALGORITHMS

Clustering is the categorising of items into multiple groups or the splitting of a data collection into subsets based on common attribute[11]. The clustering problem has been found and handled in a variety of contexts, with positive results in a variety of medical applications. Clustering medical data into tiny, meaningful groups can help with pattern discovery by allowing the extraction of a large number of acceptable features from each cluster, thereby introducing structure to the data and facilitating the use of traditional DM techniques [12].To uncover the linkages and patterns that exist in those data pieces, several similarity and dissimilarity measures are used.

3.1 KM Clustering Algorithm

The KM algorithm is a well-known clustering technique that is used in a wide range of scientific and industrial applications [13]. KM divides the data into k different clusters based on their distinctive values. The feature values of data in the same cluster are identical. The positive integer k, which represents the number of clusters, must be given ahead of time. Following are the steps involved in a KM algorithm:

The K-Means clustering approach was used to predict cardiac disease.

1. In the space, K pointers signifying the data to be clustered are placed. The primary group centroids are represented by these points.
2. The data is assigned to the neighbouring group to the centroid.
3. Once all of the data has been assigned, the positions of all of the K centroids are recalculated.
4. Repeat steps 2 and 3 until the centroids are no longer moving. As a result, data is divided into groups, from which the metric to be reduced can be deduced [14].

The KM approach with the K values is used to cluster the preprocessed heart disease data. The KM clustering algorithm produces a set of disconnected and non-hierarchical groups. It's ideal for producing globular clusters.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

3.2 Canopy Clustering(CC) Algorithms

The CC algorithm is a pre-clustering approach that is unsupervised. It's frequently used as a preprocessing step before applying the KM or Hierarchical clustering algorithms. Its purpose is to hurry up clustering processes on huge data sets where utilizing another approach directly would be unfeasible due to the data set's size [15].

The algorithm of CC works as follows: T1 (loose distance) and T2 (tight distance) are used as thresholds, with T1>T2.

1. Begin with the data points that will be grouped.
2. Remove a point from the set and use it to start a new 'canopy.'
3. Assign each remaining point in the set to the new canopy if its distance from the canopy's starting point is less than the loose distance T1.
4. If the point's distance is likewise less than the tight distance T2, remove it from the original set.
5. Continue from step 2 until there are no more data points to cluster in the set.
6. Using a more expensive but precise technique, these clustered canopies can be sub-clustered.

.

An estimated and rapid span metric could be used for step 3 as an additional speed boost, whereas a more precise and slow distance metric could be utilized for step 4.

3.3 Farthest First Techniques (FFT)

The FFT is a fast and greedy algorithm [16]. During this algorithm the primary center will be chosen haphazardly. The second center will be determined illiberally select because the point farthest from the primary.

```
1. Farthest First Traversal (D: data set, k: integer)
   {
2. randomly select first center;
3. //select centers
4. for (I= 2…, k) {
5. for (each remaining point) {calculate distance
   to the current center set;}
6. select the point with maximum distance as
   new center;}
7. //assign remaining points
8. for (each remaining point) {
9. calculate the distance to each cluster center;
10. put it to the cluster with minimum distance;}}
```

IV. EXPERIMENTAL RESULTS

The K Means, canopy clustering and FF methods are examples of clustering algorithms[17]. Using WEKA tool[18] these two algorithms were applied to a heart disease prediction dataset, and their performance was assessed. The dataset provided by the

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

American Heart Association is used in this experiment to evaluate the system's performance. There are 50 instances and ten attributes in the AHA dataset [19].

A total of twelve major Risk Factors(RF) were included in the dataset. The AHA dataset was converted to a CSV file and uploaded to the Weka programme, which used multiple clustering techniques to predict HD. The RFs and their encriped values that would be employed as input for the network, are shown in below table.

TABLE 1: DATASET OF AHA

Name	Description
Sex	Male(1), Female(0)
Age	20-34(-2),35-50(-1),51-60(0),61-79(1),>79(2)
Blood Cholesterol	Below 200 mg/dL - Low (-1), 200-239 mg/dL - Normal (0), 240 mg/dL and above - High (1)
Blood Pressure	Below 120 mm Hg- Low (-1) 120 to 139 mm Hg- Normal (0), Above 139 mm Hg- High (-1)
Hereditary	Family Member diagnosed with HD - Yes (1) Otherwise No (0)
Smoking	Yes (1) or No (0)
Alcoholic Intake	Yes (1) or No (0)
Physical Activity	Low (-1) , Normal (0) or High (1)
Diabetes	Yes (1) or No (0)
Diet	Low (-1) , Normal (0) or High (1)
Obesity	Yes (1) or No (0)
Stress	Yes (1) or No (0)
Heart Disease	Yes (1) or No (0)

Experimental results of K Means clustering techniques on AHA dataset is shown below:

Final cluster centroids:

	Cluster#		
Attribute	Full Data (50.0)	0 (21.0)	1 (29.0)
Sex	Male	Male	Female
Age	43.16	2.1429	36.6552
Blood Cholestrol	Low	High	Low
Blood Pressure	Normal	Normal	Normal
Hereditary	No	No	No
Smoking	No	No	No
Alcohol Intake	Yes	Yes	No
Physical Activity	Normal	Normal	High
Diabetes	No	Yes	No
Diet	Normal	Normal	Normal
Obesity	No	Yes	No
Stress	No	Yes	No
Heart Disease	No	Yes	No

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

KM
==========
Number of iterations: 4

Within cluster sum of squared errors:

191.42778973457285

Initial starting points (random):

Cluster-0: Female,45, Normal,High,Yes,Yes, No,

'Normal ', Yes, Normal, Yes, Yes, No

Cluster-1: Female,25,Low,Normal,No,No,

No,High,No,Poor,No,Yes,No

Time taken to build model (full training data):

0.02seconds

=== Evaluation on training set ===

Clustered Instances(CI)

0 21 (42%)
1 29 (58%)

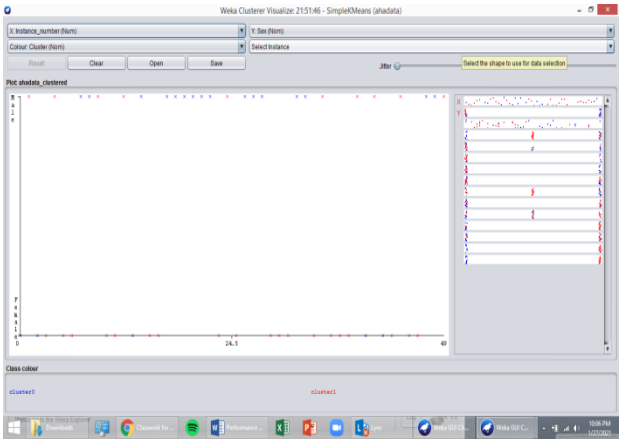

Fig 1: Result Plot of K Means Clustering

Experiment of Canopy clustering
techniques on AHA dataset is shown below:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

```
Cluster 0: Male,60,High,High,No,Yes,Yes,'Normal ',
        Yes, Normal,No,Yes,Yes,{2} <0,8>
Cluster 1: Female,31.384615,Low,Normal,No,No,No,
        High,No,Normal,No,No,No,{13} <1,4,5,6,7>
Cluster 2: Male,60.5,Normal,High,No,Yes,Yes,Low,
        Yes,Poor,Yes,No,Yes,{2} <2,10>
Cluster 3: Female,51,High,High,No,No,No,'Normal ',
        Yes,Poor,Yes,Yes,Yes,{4} <3,11>
Cluster 4: Female,37,Normal,Normal,No,No,No,'Normal ',
        Yes,Poor,No,No,No,{4} <1,4,6,7>
Cluster 5: Male,66.5,Low,Low,No,No,Yes,High,Yes,Normal,
        No,No,No,{2} <1,5,6>
Cluster 6: Female,39.666667,Normal,Normal,No,No,Yes,
        High,Yes,Normal,No,Yes,No,{3}<1,4,5,6,10>
Cluster 7: Female,42,High,Normal,No,No,No,Low,No,Poor,
        Yes,No,No,{2} <1,4,7,9,11>
Cluster 8: Male,40.5,High,Normal,No,No,Yes,'Normal ',No,
        Normal,Yes,Yes,Yes,{2} <0,8,9,10,11>
Cluster 9: Male,63.5,High,Normal,No,No,Yes,Low,No,Poor,
        Yes,Yes,Yes,{2} <7,8,9,10,11>
Cluster 10: Male,69.5,Normal,Normal,No,No,Yes,Low,Yes,
        Normal,Yes,Yes,Yes,{2} <2,6,8,9,10,11>
Cluster 11: Female,40,High,Normal,No,No,Yes,Low,Yes,
        Poor, Yes,Yes,Yes,{2} <3,7,8,9,10,11>
```

CC

==========

Number of canopies (cluster centers) found: 12
T2 radius: 1.771
T1 radius: 2.214

Time taken to create model (full training data):
0.03 seconds
=== Evaluation on training set ===

CI

0	5 (10%)
1	12 (24%)
2	2 (4%)
3	7 (14%)
4	4 (8%)
5	3 (6%)
6	4 (8%)
7	3 (6%)
8	4 (8%)
9	2 (4%)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

10 2 (4%)
11 2 (4%)

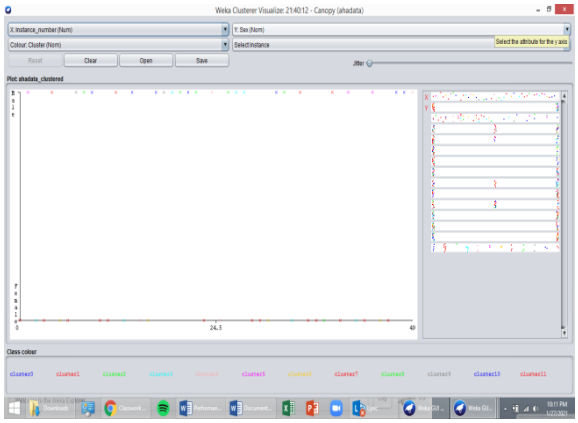

Fig2: Result Plot of Canopy Clustering

Experiment of Farthest First clustering techniques on AHA dataset is shown below:

Farthest First

==========

Cluster centroids:

Cluster 0 Male 42.0 Low Normal No No

Yes Low No Poor No No No

Cluster 1 Female 45.0 Normal High Yes Yes

No Normal Yes Normal Yes Yes No

Time taken to create model (full training data): 0.02 seconds

=== Evaluation on training set ===

CI

0 34 (68%)

1 16 (32%)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1363 - 1371
https://publishoa.com
ISSN: 1309-3452

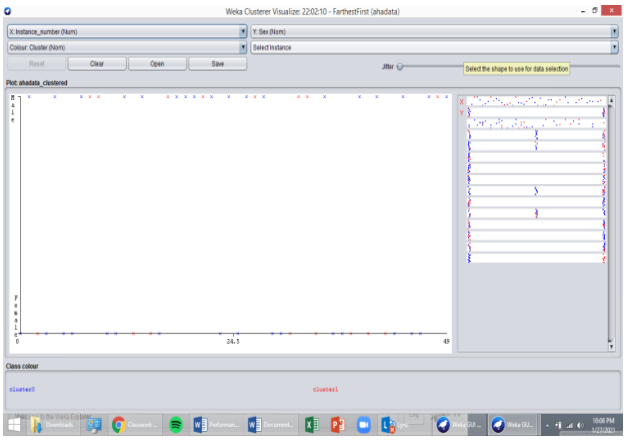

Fig 3: Result Plot of FFT Clustering

The Execution of KM ,CC and FFT were assessed using the following variables.

TABLE 2: EXECUTION OF THREE CLUSTERING TECHNIQUES

Clustering Techniques used	No. of clusters	Time taken
K M	2	0.02 seconds
CC	11	0.03 Seconds
FFT	2	0.02 Seconds

The AHA dataset is run via the Weka tool, the number of clusters and time spent for the KM, CC and FFT were recorded.

V. CONCLUSIONS

The main intention of this study is to forecast diseases based on medical data sets. The major goal of this research is analyzing different clustering approaches in order to find the most suited data mining technique for predicting HD at an early stage. The AHA dataset is used in this work, which is subjected to several clustering techniques utilizing Weka. Clustering is a type of observation-based unsupervised learning technique. Clustering analysis' major purpose is to group both comparable and dissimilar items in the same clusters and separate clusters, respectively. The KM, CC and FFT approaches were evaluated in terms of the number of clusters and time taken for execution. This work demonstrated how to obtain clusters and calculate the new centroid for high-dimensional datasets using K-means, CC and FFT. These clusters, which were created using three different algorithms, can then be utilized as input into classification to acquire the best accuracy for HD prediction.

REFERENCES

[1] World Health Organization. (2016). Hearts: technical package for cardiovascular disease management in primary health care.
[2] Devi, M. R. (2016). Analysis of various data mining techniques to predict diabetes mellitus, International Journal of Applied Engineering Research, 11(1). 727-730
[3] Reetu Singh and E.Rajesh," Prediction of Heart Disease by Clustering and Classification Techniques" International Journal of Computer Sciences and Engineering Open Access Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

[4] Arun K. Pujari, ―Data Mining Techniques‖, Universities Press (India) Ltd, 2001.
[5] S. Vijayarani and S. Sudha," An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples , Indian Journal of Science and Technology, Vol 8(17), DOI: 10.17485/ijst/2015/v8i17/52123, August 2015
[6] Jonathan.C.Prather, M.S. "Medical Data Mining: Knowledge Discovery in a clinical Data warehouse ", 1995.

[7] Madhulatha TS. An overview on clustering methods. IOSR J Eng. 2012;2(4):719–725.
[8]Alkadhwi Ali Hussein Oleiwi and AdelajaOluwaseunAdebayo ,"Data Mining Application Using Clustering Techniques (K-Means Algorithm)In The Analysis Of Student's Result", Journal of Multidisciplinary Engineering Science Studies (JMESS),ISSN: 2458-925X,Vol. 5 Issue 5, May – 2019
[9] Bala Sundar V, Devi .T and Saravanan.N "Development of a Data Clustering Algorithm for Predicting Heart",International Journal of Computer Applications (0975 – 888),Volume 48– No.7, June 2012
[10] McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123

[11] Zakaria Nouir, Berna Sayrac, Benoît Fourestié, Walid Tabbara, and Françoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications,Vol. 2, No. 6, pp. 30-37, November 2007.
[12] F. H. Saad, B. de la Iglesia, and G. D. Bell, ― A Comparison of Two Document Clustering Approaches for Clustering Medical Documents, Proceedings of the 2006 International Conference on Data Mining (DMIN-06), 2006.
[13] C. Ordonez, ― Programming the K-Means Clustering Algorithm in SQL, Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.
[14] Shantakumar, B. Patil and Y.S Kumaraswamy.,―Intelligent and Effective Heart attack Prediction System Using Data Mining and Artificial Neural Network, Eurp Journals Publishing Inc. ISSN 1450-216X Vol.31 No.4 2009, pp.642-656, 2009.
[15]https://en.wikipedia.org/wiki/Canopy_clusteringalgorithm#cite_note-original-1
[16] [Mr. Rinal H. Doshi Dr. Harshad B. Bhadka Ms. Richa Mehta" Development Of Pattern Knowledge Discovery Framework Using Clustering Data Mining Algorithm" International Journal of Computer Engineering & Technology (Ijcet) Volume 4, Issue 3, May-June (2013), Pp. 101-112.
[17] Dr.V. Poornima andSarala Devi. U, "Performance evaluation of triumvirate clustering algorithms for heart disease prediction" ,European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 11, Pages 7780-7789
[18] Data mining in bioinformatics using Weka.Frank E, Hall M, Trigg L, Holmes G, Witten IH Bioinformatics. 2004 Oct 12; 20(15):2479-81
[19] American Heart Association. (2014). Atherosclerosis http://www.heart.org/HEARTORG/Conditions/Cholesterol/WhyCholesterolMatters. Atherosclerosis_UCM_305564_Article. jsp (18 Maret 2014).