Heart ailments prediction using Data Mining Techniques

[1]Priyanshi Burad, [2]Manik Gautam

Computer Science
Mukesh Patel School of Technology, Management and Engineering
Shirpur, India

Abstract: Errors in the field of medical are quite risky for human beings which results in the increase in the number of casualties. Cardiovascular disease prediction has become very relevant in the modern day. The author has reviewed 33 papers related to heart disease prediction involving different methods and algorithms. Human errors are possible and to be safe from such errors, many techniques are developed to predict heart diseases. Here we will compare amongst the techniques which will better serve our purpose for predicting heart disease more accurately.

Keywords: **Data Mining, Heart disease prediction, Classification algorithms.**

I. INTRODUCTION

With this growing age of technology people are mostly involved with their monotonous life and ill habits. The ill habits can include egregious food habits, smoking, alcohol consumption and not exercising regularly. These habits can be termed as risk factors and they have a direct impact on our heart. Fatalities due to heart diseases in India increased from 1.3 million to 2.8 million between the years 1990 and 2016. Every 60 seconds a person dies of heart related disease in India [34]. These numbers will keep on rising in the future.

The techniques used by doctors earlier have been erroneous, geriatric and obsolete. The cases of false positive and false negative are also sky rocketing to the sky. With the advancement of the techniques adopted by the medical practitioners it has now become easy to predict the heart diseases with the certain factors like age, gender, previous medical record, family history and so on with a very good amount of accuracy.

Data mining and predictive analysis methods are largely employed in the medical fields to easily detect the disease up to the maximum accuracy. Data mining is the branch of computers which is based on the extraction of useful and meaningful data from the raw data sets. Data mining and predictive modelling have been an asset to the medical fields. There are quite a number of data-mining algorithms that has been listed in this review paper. Bayes classification, apriori-algorithm, decision tree and Radial BasisFunction (RBF) are to name a few. Every algorithm has different precision and accuracy. These algorithms are usually supported by various software to give us more detailed and finer results.

II. DATA MINING TECHNIQUES

A. Decision Tree

It is used in classification, and prediction tasks. It uses a divide and conquer approach to solve a problem into small subsets. It is an inductive approach. The technique consists of roots, nodes, branches and leafs. Here, test on attribute is represented by the internal nodes of the tree, the branch shows the outcome of the test and class is represented by the leaf.

Figure I. Decision Tree.

B. Naive Bayes

It is the classification algorithm works on the concept of bayes rule of conditional probability. If a problem instance is given to be classified, it is written as x=(x1 ,….., xn) representing some dependent variables, which is assigned to instance probability. p(Ck | x1,…...,xn) for each possible classes or k outcomes.

C. SVM

This machine learning algorithm is mostly used for classification problems and issues. Support vectors are the coordinates in the x-y plane which represent the individual observations and support vector machine distinguishes or separates among the classes using a hyper plane.

D. ANN

It acts as a directed graph which consists of input, output and hidden nodes. They are also names as source, sink, and internal nodes. The neural network can learn from the previous data and then they can modify themselves. The basic functionality of the ANN is that it can work like a human brain.

E. KNN

K-nearest neighbor is used in both regression as well as in classification, this algorithm reduces calculations and interpretation becomes simple. It is a supervised learning method, its performance is also good even if the data set is large and have some noise in it. To identify the class for a new set, we must have the information regarding the distance of it from other items in the training set.

F. Maximum Frequent Item set Algorithm (MAFIA)

This algorithm is efficient for using a dataset which is huge and vast. Here, some of the dataset's part is cropped or pruned and is compressed in such a way, that we are left with minimum data that is needed to be assessed.
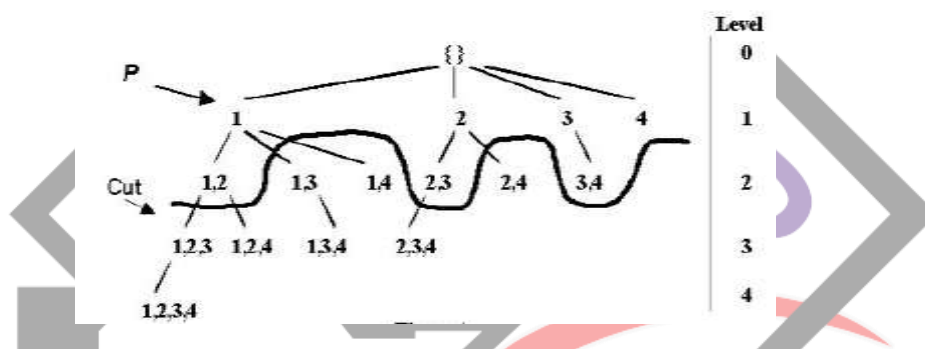

Figure II. Maximum Frequent Item set Algorithm

G. Apriori

It is an iterative approach where we can find k+1 data set from k frequent datasets. Apriori applies level-wise search approach and is a tree-based algorithm. It scans data multiple times to generate candidate key. We can improve the efficiency of Apriori by using techniques based on hash, reduction of transaction, partitioning, dynamic dataset count and sampling.

H. K means

This algorithm comes under unsupervised learning which is used to solve clustering problems iteratively. In this technique we have to deal with unlabeled data and that data is clustered based on their similar features. The term "means" in the algorithm refers to finding the average in data, that is to find centroid and this centroid is used to label new data. Here, data which is similar are clustered together.

I. K.MLP

MLP is a feed forward neural network in which neurons are placed inside the layers with the output always flowing towards the output layer. If there is existence of only one layer then it's called perceptron and if multilayer then it is called multi-layer perceptron basically it is a network of perceptron.

J. Random Forest

Random Forest algorithm is the combination of decision tree (cart model) and bootstrapping algorithm. Different samples and different initial variables are used to build up cart model. This process of building Cart model is repeated and then final prediction about individual observations is made. The prediction which is final a function of each individual prediction. The prediction which is final can be calculated by taking the mean of each individual prediction.

K. Simple Logistic Regression

This technique works with binary data where 0 denotes the failure and 1 denotes its success. Logistic regression uses logarithmic functions and finds relationship between variables, and predicts the future data points from that function. Its main goal is to find the probability of occurrence of event.
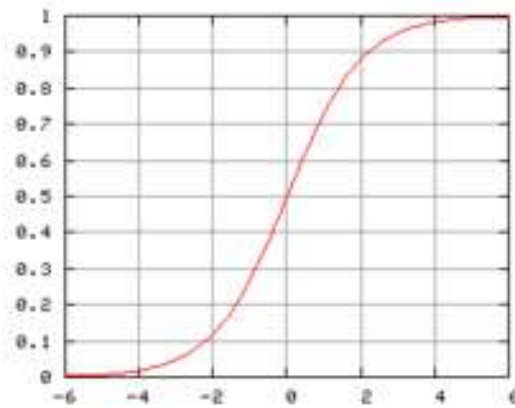

Figure III. Simple Logistic Regression

L. Radial Basis Function (RBF)

Any function is known as radial function if it satisfies the propertywhere the value of the function depends upon its distance from its origin. They are used as a kernel in SVM. They also have three layers similar to neural networks that are hidden layer, input layer and output layer.

M. Farthest First

It is a technique where an arbitrary point is selected and each point is placed far from the selected arbitrary points.

III. LITERATURE REVIEW

The authors have analyzed about 33 papers in which the most common attributes pertaining to heart disease prediction are chest pain, sex, age, blood pressure, cholesterol, fasting blood sugar, electro cardio-graphic result, maximum-heart rate, Exang, Oldpeak, Slope, Ca, Thal.

Table I. Description of 29 papers individually.

Sr no.	Research (Year)	Technique implemented	Findings
1	TanmayTamhane *et al.*, 2015 [1]	Naive Bayesian Algorithm	14 attributes are used where the predicted accuracy if a person is having heart disease are: Naive Bayesian Algorithm : 74.2% Weighted Associative Classifier : 74.2%
		Weighted Associative Classifier	
2	Saeidehkabirirad *et al.*, 2016 [2]	Multilayer perceptron (MLP)	Total number of samples is 1215 and attributes are 13. Accuracy achieved by these algorithms are: Multilayer perceptron : 98% RBF : 95%
		RBF	
3	MeenuSingla *et al.*, 2016 [3]	K-Means	Pima diabetes dataset which contains 768 instances and 9 attributes was used. Out of these techniques which are used in this paper, the Farthest First technique is best as it requires less execution time that is of 0.02 seconds and the ratio of each cluster is 80:20 which is maximum if compared to the other techniques stated.
		EM Technique	
		Farthest First	

4	Aditya Jain *et al.*, 2018 [4]	Weighted Association Rule	They have used KEEL software and in this they got the accuracy of 61.2% in training dataset and 53.1% in test data set.
5	ChaitraliS *et al*, 2012 [5]	Neural Networks	Data set of 573 observations and 13 attributes was used. Training dataset consists of 303 records and Testing dataset consists of 270 records. The accuracy obtained by the techniques used in this paper are: Neural Networks: 100% Decision-Tree: 99.62 % Naïve Bayes:90.74%
		Decision Tree	
		Naive Bayes	
6	AditiGavhane *et al.*, 2017 [6]	neural network algorithm multi-layer perceptron (MLP)	They have used 7 attributes. Here as an output we will get prophecy result in terms of Yes or No if the person has heart disease. Its precision is 91%.
7	KanakSaxena *et al.*, 2016 [7]	Rules generated by the authors: Original, Pruned, without duplicates, Classified, Sorted, Polish	Accuracy of 86.7% is achieved by using 13 parameters.
8	Marimuthu *et al.*, 2018 [8]	SVM	SVM : 83.60 Naïve Bayes: 80.66 Decision Tree : 75.58 KNN : 65.56
		K-nearest	
		Naive Bayes	
		Decision Tree	
9	Mohd Ashraf *et al.*, 2019 [9]	Deep Neural Network	The accuracy achieved is 87.64%
10	Hlaudi Daniel *et al.*, 2014 [10]	J48	12 attributes were used and the accuracy results are: J48 :99.0741% REPTREE:99.0741 Naïve-Bayes :97.222% Bayes S Net:98.1481 % Simple Cart : 99.0741%
		REPTREE	
		Naive Bayes	
		Bayes S Net	
		Simple Cart	
11	M. Anbarsi *et al.*, 2010 [11]	Naive Bayes	13 attributes have been reduced to 6 with the help of genetic search and maintained the accuracy of algorithms. Naïve-Bayes: 96.5% Decision-Tree: 99.2% Classification-via clustering :88.3%
		Decision Tree	
		Classification via clustering	
12	ShrutiRatnaka *et al.*, 2013 [12]	Naive Bayes	Naïve-Bayes :96.5% Decision-Tree :99.2%
		Decision Tree	
13	JyotiSoni *et al.*, 2011 [13]	Weighted Associative Classifiers	303 records with 14 attributes and the accuracy was 81.51%.
14	Ria A Kurian *et al.*, 2018 [14]	KNN	KNN : 81.2% Decision Tree :79.06% NaiveBayes : 86.4%
		Decision Tree	

		Naïve Bayes	Ensemble Classifier : 90.8%
		Ensemble Classifier	
15	ResulDasa *et al.*, 2009 [15]	Neural network block	SAS enterprise miner was used as a tool which resulted in 89.01 % classification accuracy.
16	S.SHARMILA *et al.*,2017 [16]	Improved Naive Bayes	Dataset of 303 records with 14 risk factors was implemented. The proposed method showed the accuracy of 97%, which was 8% more than naive bayes.
17	Aditya Methaila *et al.*,2014 [17]	Apriori	15 attributes were used and accuracy achieved are: Decision Tree: 99.62% Naive bayes-96.53
		decision tree	
		naive bayes	
		mafia algorithm	
18	Navdeep Singh *et al.*,2019 [18]	Hybrid Genetic Naive Bayes(proposed method) was used compared with existing models like Logistics Regression, Naïve Bayes, Attribute Weighted Artificial Immune System, Modified Artificial Immune System, Neural Networks ensemble, ANN_Fuzzy_HP	using the proposed method of hybrid genetic naive bayes we got the accuracy of 97.14%
19	H. Benjamin Fredrick David *et al.*,2018 [19]	Random forest	Accuracy of 92% is achieved by random forest algorithm if compared with Decision-tree algorithm and Naïve-Bayes.
		naive bayes	
		decision trees	
20	K. Subhadra *et al.*,2019 [20]	multilayer perceptron neural network with back propagation	Accuracy and Precision of 94% and 90% respectively
21	Kathleen H. Miao *et al.*,2018 [21]	Deep neural network	accuracy and precision of 83.67 and 79.12 respectively
22	Dr. K. Aravinthan *et al.*,2018 [22]	proposed method is a mixture of clustering techniques	Accuracy of 98% is achieved.
23	M.Akhiljabbar *et al.*,2012 [23]	Apriori algorithm with hypothesis testing (Z Testing)	The proposed method had an accuracy of 88.9%
24	Mai Shouman *et al.*,2012 [24]	K-nearest neighbour	KNN achieved an accuracy of 97.4
25	A. Durga Devi *et al.*,2015 [25]	Radial basis function (RBF) network	The proposed system gave an accuracy of 83.83% with 14 attributes and 85.48 with 5 attributes.
26	Harsh Vazirani *et al.*,2010 [26]	Modular neural network	The accuracy was 87.02% in probabilistic sum integration and 89.72% in probabilistic product integration.

27	R.Chitra *et al.*,2013 [27]	CNN	Accuracy achieved are: CNN (Training set):78.55% CNN (Testing dataset): 85%
		SVM	SVM (Training set):75%% SVM (Testing dataset):82%
28	TS. Spino *et al.*, 2019 [29]	Naive Bayes	The accuracy of Naive Bayes is 98.8636%.
29	Ebenezer ObaloluwaOlaniyi *et al.* ,2015 [30]	Back propagation Neural	BPNN with an accuracy of 85% and SVM with an accuracy of 87.5%.
		Network (BPNN)	
		Support Vector Machine	

C.Sowmiya et al., used apiori algorithm and SVM for diagnosis of heart disease and they even proved that classification techniques are efficient and gives high accuracy.[30]

Nitit Guru et al., have implemented heart disease diagnosis using Neural networks by using 13 attributes and database containing 78 patients record, they concluded that neural network has huge scope in medical science.[31]

SarathBabu et al., implementation of K-means algorithms, MAFIA algorithm and Decision tree classification has been observed and after working with 14 attributes decision tree performed better amongst the techniques stated. [32]

S. Kiruthika Devi et al., they have increased the accuracy of the predicted outcome by using a hybrid of different techniques and implemented the dataset on software like weka, orange and Rapid Miner.[33]

IV.　RESULTS AND DISCUSSION

It has been observed that in [2], [5], [10], [11], [12], [16], [17], [18], [22], [24], [28] we have achieved accuracy above 95%. RBF, MLP, Decision Tree, Naïve Bayes has performed tremendously well. It is observed that decision tree has outperformed in category of classification techniques as we can see in [5],[8],[11],[12].

V.　CONCLUSION

The accuracy of different data mining techniques has been mentioned above. It is observed that accuracy and precision of each data-mining technique is subject to the number of attributes and the tools used. It is highly recommended that the algorithm that is needed for the further research work must depend on the attributes. All the above mentioned techniques have a fair amount of accuracy.

REFERENCES

[1] TanmayTamhane, Mateen Shaikh, SanjaykumarBoga, MrunalTanwar, A.E. Patil "Heart Disease and Diabetes Prediction using Data Mining *" in IJSRD - International Journal for Scientific Research & Development| Vol. 3, Issue 01, 2015 | ISSN (online): 2321-0613*

[2] SaeidehKabrirada, HosseinKardanmoghaddamb, VahidrezaAfshinc a "Heart Disease prediction by using artificial neural network" *in International Journal of Computer Science and Information Security (IJCSIS), Vol 14, No. 1, Janurary 2016*

[3] MeenuSingla, KawaljeetSingh, "Heart Disease Prediction System using Data Mining Clustering Technique "inInternational Journal of Computer Applications (0975 – 8887) International Conference on Advances in Emerging Technology (ICAET 2016)

[4] Aditya Jain, Purushottam Sharma, Vikas Deep, Aakash Chauhan, "Heart Disease Prediction using Evolutionary Rule Learning"*in International Conference onComputational Intelligence and Communication Technology" (CICT 2018)*

[5] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Technique" *in International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012*

[6] AditiGavhane, Isha Pandya, GouthamiKokkula, Prof. Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning" inProceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1

[7] Prof.(Dr).KanakSaxena, Purshottam, Richa Sharma, "Efficient Heart Disease Prediction System" *in Procedia Computer Science 85 (2016) 962 – 969*

[8] Marimuthu, S.Deivarani, Gayathri.R, "Analysis of Heart Disease Prediction using Various" in Conference: International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, (AISGSC 2019), At PSG College of Technology, Coimbatore - 641 004

[9] Mohd Ashraf, M. A. Rizvi, Himanshu Sharma, "Improved Heart Disease Prediction Using Deep Neural Network*" in Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.2, 2019, pp. 49-54*

[10] Hlaudi Daniel Masethe, Mosima Anna Masethe " Prediction of Heart Disease using Classification Algorithms "in Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA

[11] M. ANBARASI , E. ANUPRIYA, N.CH.S.N.IYENGAR " Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm" in International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[12] SHRUTI RATNAKA, K. RAJESWARI, ROSE JACOB "Prediction of heart disease using Genetic Algorithm for selection of optimal reduced set of attribute" in international Journal of Advanced Computational Engineering and Networking, ISSN (p): 2320-2106, Volume-1, Issue-2, April-2013

[13] JyotiSoni, Uzma Ansari, Dipesh Sharma, SunitaSoni "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers" in International Journal on Computer Science and Engineering (IJCSE)

[14] Ria A Kurian, Lakshmi K.S "An Ensemble Classifier for the Prediction of Heart Disease" in International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 6| ISSN : 2456-3307

[15] ResulDasa, Ibrahim Turkoglub, AbdulkadirSengurb "Effective diagnosis of heart disease through neural networks ensembles" in Expert Systems with Applications 36 (2009) 7675–7680

[16] S.SHARMILA, Dr.M.P.INDRA GANDHI "Analysis of Heart Disease Prediction Using Data Mining Techniques" in International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 93-95 (2017) Special Issue

[17] Aditya Methaila , Prince Kansal , Himanshu Arya , Pankaj Kumar "Early heart disease prediction using data mining techniques" in CS & IT-CSCP 2014 10.5121/csit.2014.4807

[18] NavdeepSingh,Sonika Jindal "Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms " in International Journal of Advance Research, Ideas and Innovations in Technology

[19] H. Benjamin Fredrick David , S. Antony Belcy "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES " in ICTACT JOURNAL ON SOFT COMPUTING, OCTOBER 2018, VOLUME: 09, ISSUE: 01

[20] K. Subhadra, VikasB "Neural Network Based Intelligent System for Predicting Heart Disease" in International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-5 March, 2019

[21] Kathleen H. Miaoa , Julia H. Miaoa" Coronary Heart Disease Diagnosis using Deep Neural Networks " in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 10, 2018

[22] Dr. K. Aravinthan "Heart Attack Prediction Using Data Mining Techniques " in International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018, 16119-16123

[23] M.Akhiljabbar , Dr.PritiChandrab , Dr.B.LDeekshatuluc "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" in International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

[24] Mai Shouman, Tim Turner, and Rob Stocker " Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients" in International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[25] A. Durga Devi "Enhanced Prediction of Heart Disease by Genetic Algorithm and RBF Network" in International Journal of Advanced Information in EngineeringTechnology (IJAIET) ISSN: xxxx: Vol.2, No.2, February 2015

[26] H. Vazirani, R. Kala, A. Shukla, R. Tiwari "Use of Modular Neural Network for Heart Disease" in International Journal of Computer and Communication Technology 1(2-4), 88-93

[27] R. Chitra and Dr.V. Seenivasagam"Heart Disease Prediction System Using Supervised Learning Classifier" in Bonfring International Journal of Software Engineering and Soft Computing, Vol. 3, No. 1, March 2013

[28] S. Spino, Dr. M. Mohamed Sathik, Dr. S. ShajunNisha "THE PREDICTION OF HEART DISEASE USING NAIVE BAYES CLASSIFIER" in International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 03 | Mar 2019

[29] Ebenezer ObaloluwaOlaniyi ,OyebadeKayodeOyedotun ,Khashman Adnan " Heart Diseases Diagnosis Using Neural Networks Arbitration" in I.J. Intelligent Systems and Applications, 2015, 12, 72-79 Published Online November 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2015.12.08

[30] C.Sowmiya, Dr.P.Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques" in 2017 IEEEINTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL,OPTIMIZATION AND SIGNAL PROCESSING

[31] NitiGuru,AnilDahiya, NavinRajpal "Decision Support System for Heart Disease Diagnosis Using Neural Network" in Delhi Business Review Vol. 8, No. 1 (January - June 2007)

[32] SarathBabu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique" in International Conference on Electronics, Communication and Aerospace Technology ICECA 2017

[33] S. Kiruthika Devi*, S. Krishnapriya and DristiponaKalita "Prediction of Heart Disease using Data Mining Techniques" in Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016

[34] https://www.business-standard.com/article/health/15-of-deaths-in-india-were-due-to-heart-diseases-in-1990-now-up-to-28-118091800130_1.html