Heart Disease Prediction Using Effective Machine Learning Techniques

Avinash Golande, Pavan Kumar T

ABSTRACT---In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and female category and this ratio may vary according to the region also this ratio is considered for the people of age group 25-69. This does not indicate that the people with other age group will not be affected by heart diseases. This problem may start in early age group also and predict the cause and disease is a major challenge nowadays. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases.

Keywords— Classification, Heart Disease, Decision Tree, Data Mining.

I. INTRODUCTION

The contents of this paper mainly focus on various data mining practices that are valuable in heart disease forecast with the assistance of dissimilar data mining tools that are accessible. If the heart doesn't function properly, this will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. In today's era heart disease is the primary reason for deaths. WHO-World Health Organization has anticipated that 12 million people die every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is a sort of heart disease that occurs due to strengthening, blocking or lessening of blood vessels which drive through the brain or it can also be initiated by high blood pressure [1].

The major challenge that the Healthcare industry faces now-a-days is superiority of facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted. [2]

Records or data of medical historyis very large, but these are from many dissimilarfoundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database.

Even if cardiovascular diseases is found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on on the well-timed judgment of that disease. An correct and methodical tool for recognizing high-risk patients and mining data for timely analysis of heart infection looks a serious want.

Different person body can show different symptoms of heart disease which may vary accordingly. Though, they frequently include back pain,jaw pain, neck pain, stomach disorders,and tininess of breath,chest pain, arms and shoulders pains. There are a variety of different heart diseases which includes heart failure and stroke and coronary artery disease [3].

Even though heart disease is acknowledged as the supreme chronic sort of diseasein the world, it can be most avoidable one also at the same time. A healthy way of life (main prevention) and timely analysis (inferior prevention) are the two major origins of heart disease director. Conducting steady check-ups (inferior prevention) showsoutstanding role in the judgment and early prevention of heart disease difficulties. Several tests comprising of angiography, chest X-rays, echocardiography and exercise tolerance test support to this significant issue. Nevertheless, these tests are expensive and involve availability of accurate medical equipment.

Heart expert's create a good and huge record of patient's database and store them. It also delivers a great prospect for mining a valued knowledge from such sort of datasets.

There is huge research going on to determine heart disease risk factors in different patients, different researchers are using various statistical approaches and numerous programs of data mining approaches. Statistical analysis have acknowledged the count of risk factors for heart diseases counting smoking, age, blood pressure, diabetes,total cholesterol, and hypertension, heart disease training in family, obesity and lack of exercise. For prevention and healthcare of patients who are about to have addicted of heart disease it is very important to have awareness of heart diseases.

Researchers make use of several data mining techniques that are accessible to help the specialists or physicians identify the heart disease. Commonly used procedures used are decision tree, k-nearest and Naïve Bayes. Other different classification based techniques used are bagging algorithm, kernel density, sequential minimal optimization and neural networks, straight Kernel self-organizing map and SVM (Support Vector Machine). The next section clearly provides details of techniques that were used in the study.

Revised Manuscript Received on June 10, 2019.

Avinash Golande, Research Scholar, K L University, Vijayawada, Andra Pradesh(Email- avinash.golande@gmail.com)

Dr. Pavan Kumar T, Professor, K L University, Vijayawada, Andra Pradesh(Email- pavankumar_ist@kluniversity.in)

The diseases that come under cardiovascular disease are coronary heart disease (CHD),cerebrovascular disease (Stroke), congenital heart disease, provocative heart diseases,Hypertensive heart diseases, and exterior artery disease. Among them, the tobacco chewing, unhealthy diet, physical inactivity and alcohol are the primary cause of heart diseases. Researchers are using a variety of classes of mathematical data mining tools that are existing in the study of heart diseases [4].

In our paper further we have studied various algorithms and tools which are used in identifying patients who are about to be affected by heart disease.

II. LITERATURE SURVEY

In the above study we will see different data mining techniquesthat were used to classify the heart diseases.

In year 2000, research conducted by ShusakuTsumoto [5] says that as we human beings are unable to arrange data if it is huge in size we should use the data mining techniques that are available for finding different patterns from the available huge database and can be used again for clinical research and perform various operations on it.

Y. Alp Aslandogan, et. al. (2004), worked on three different classifiers called K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and used Dempsters' rule for this three viewpoint to appear as one concluding decision. This classification based on the combined idea show increased accuracy [6].

Carlos Ordonez (2004), Assessed the problematic to recognize and forecast the rule of relationship for the heart disease. Adataset involving medical history of the patients having heart disease with the aspects of risk factors was accessed by him, measurements of narrowed artery and heart perfusion. All these restrictions were announced to shrink the digit of designs, these are as follows:

1) The features should seem on a single side of the rule.
2) The rule should distinct variousfeatures into the different groups.
3) The count of featuresavailable from the rule is organized by medical history of people having heart disease only. The occurrence or the nonappearance of heart disease was predicted by the author in four heart veins with the two clusters of rules [7].

Franck Le Duff (2004), worked on creating Decision tree quickly with clinical data of the physician or service. He suggested few data mining techniques which can help cardiologists in the predication survival of patients. The main drawback of the system was that the user needs to have knowledge of the techniques and we should collect sufficient data for creating an suitable model [8].

Boleslaw Szymanski, et. al. (2006), operated on a novel experiential to check the aptitude of calculation of scarce kernel in SUPANOVA. The author used this technique on a standard boston housing market dataset for discovering heart diseases, measurement of heart activities and prediction of heart diseases were found 83.7% correct which were measured with the help of support vector machine and kernel equivalent to it. A quality result is gained by spline kernel with the help of standard boston housing market database [9].

Kiyong Noh, et. al. (2006) made use of a classification technique for removal of multi-parametric structures by accessing HRV and ECG signals. Kiyong used the FP-growth algorithm as the foundation of this technique that is associative. A rule consistency degree was gained which allows a robust press on trimming designs in the method of producing designs[10].

HeonGyu Lee, et. al. (2007), operated for the operation systems of Arithmetical and cataloguing for the addition chief of the multi-parametric feature through direct and nonlinear features of Heart Rate Variability (HRV). The dissimilar classifiers existing are cataloguing grounded on Decision Tree (C4.5), Multiple Association Rules (CMAR) and Bayesian classifiers, and Support Vector Machine (SVM) that are investigated for the valuation of the linear and nonlinear features of the HRV tables [11].

Niti Guru, et. al. (2007), functioned for forecasting of heart disease, Blood Stress and Sugar by the aid of neural systems. Hearings were accepted out on example best ever of patients. The neural system is verified with 13 types, as blood pressure,period, angiography etc. [12].

Controlled network was used for analysis of heart diseases. Training was accepted out with the support of a back-propagation technique. The secretive data was nourished at certain times by the doctor; the acknowledged technique applied on the unidentified data since the judgments with trained data and caused a grade of possible ailments that the patient is inclining to heart disease.

Hai Wang, et al. (2008), deliberated the part of medicinal experts in medical data mining also onobtaining a model for medical awareness achievement using data mining [13].

SellappanPalaniappan, et. al. (2008), industrialized IHDPS-Intelligent Heart Disease Prediction System by means of data mining algorithm, i.e. Naïve Bayes, Decision Trees and Neural Network. Each process has its own authority to advance right results. The unknown designs and association amongst them have were used to paradigm this method. The IHDPS is web-based,user-friendly, mountable, trustworthy and stretchy and justifiable [14].

LathaParthiban, et. al. (2008), operated on the foundation of CANFIS(co-active neuro-fuzzy implication method) for identification of heart disease. CANFIS model established the disease by integrating the neural network and fuzzy logic methods and later combined with the genetic algorithm. On the grounds of the training presentations and classification correctness found, the performance of the CANFIS model were estimated. The CANFIS prototypical is exposed as the possible for estimation of heart disease [15].

Chaitrali S. D., (2012), investigated a computation structures for heart syndrome with the help of full amount of input characteristics. A few terms related to medical like blood pressure, sex, cholesterol and 13 more attributes like this were recycled to predict the heart disease to a particular person or patient. He also made use of two different attributes like smoking and obesity. Unlike data mining performances were used like Decision trees, neural networks and naïve baye's for analyzing the heart disease database. The concert of these practices depends on the accuracy provided by the system. The accuracy provided by decision tree is 99.62%, neural network is 100% and naïve bayes is 90.74% respectively[16].

S. Vijayarani, et. al. in (2013), made use of experimental results carried out using dissimilar classification methods for heart disease dataset. The different classification systems which were used and tested by him are Decision Stump,Random Forest and LMT tree algorithm. WEKA tool was used for comparison [17].

Harsh Vazirani, et. al. (2010), deliberates many belongings connected to the analysis of the heart disease. The main emphasis is on two kinds of the analysis methods that were used one is manual and the other is programmed analysis which contains of analysis of diseases with the help of segmental neural network and intellectual expert structure that were used for analyzing heart diseases [18].

Different characteristics are separated, agreed to the two neural network algorithms i.e. Back-propagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFNN) for schooling and challenging.

Accuracy obtained using different techniques:

Writer	Year	Methods/Techniques	Count ofattributes
Carlos et. al.	2001	Association Rule Mining	25
Latha et. al.	2008	Genetic Algorithm	14
		CANFIS	
Shantakumar et. al.	2009	MAFIA	13
		Clustering	
		K-Means	
Dr. K. Usha Rani	2011	Classification	13
		Neural Network	
Majabbar, et. al.	2011	Clustering	14
		Association Rule Mining	
		Sequence Number	
Nan-Chen, et. al.	2012	(EVAR)	
		Machine Learning	
		Markov Blanket	
Oleg, et. al.	2012	Artificial Neural Network	
		Genetic Polymorphisms	
Shadab, et. al.	2012	Navie Bayes	15
NidhiBhatla, et. al.	2012	Fuzzy Logic	4
		Weka Tool	
		Decision Tree	
		Naive Bayes	
		Classification via. Clustering	
JesminNahar,et. al. Ms. Ishtake, et. al.	2013 2013	AprioriPredictive AprioriTertius Decision Tree	14 15
		Neural Networks	
		Navie Bayes	
Ashish Kumar Sen1, et. al.	2013	Neuro-fuzzy	4
		Backpropagation Algorithm	
KanteshKumar Oad et. al.	2014	Fuzzy Rule Based Support System	6

Year	Author	Purpose	Techniques used	Accuracy
2013	AbhishekTanej a	Heart disease prediction system using data mining techniques and different supervised Machine learning algorithms	J48	95.56%
			SMO	92.42%
			Multilayer perception	94.85%
2015	Acharya et al.	diabetic subject by heart rate variability signals	KNN	Highest accuracy of 92.02% obtained by DT

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

			Naïve Bayes (NB)	
			SVM	
2015	Priti Chandra et al.	Computational Intelligence Technique for early	Naïve Bayes	86.29%
2015	Cemil et al.	Propose application of knowledge discovering process on prediction of stroke patients	ANN	81.82% for training dataset
				85.9% for test data set
			SVM	80.38% for train data set
				84.26% for test data set
2016	Muhammad Saqlain et al.	Identification of Heart Failure by Using Unstructured	Logistic Regression	80.00%
			Neural Network	84.80%
			SVM	83.80%
			Random Forest	86.60%
			Decision Tree	86.60%
			Naïve Bayes	87.70%
2016	Marjia et al.	Heart disease prediction using WEKA tool and 10-Fold cross-validation	KStar	75%
			J48	86%
			SMO	89%
			Bayes Net	87%
			Multilayer Perceptron	86%
2016	Dr. S. Seema et al.	Predict chronic disease by mining the data containing in historical health records	Naïve Bayes	Highest accuracy in case of heart disease 95.556% is achieved by SVM.
			Decision tree Support Vector Machine (SVM)	
2016	Tapas RanjanBaithar u	Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset	J48	68.97%
			ZeroR	57.97%
			Multilayer Perceptron	71.59%
			IBK	62.90%
			Naïve Bayes	55.36%
			VFI	60.29%
2016	VidyaK.Sudarshan et al.	Application of higher-order spectra for the characterization of coronary artery disease using electrocardiogram signals.	KNN	98.17%
			Decision Tree (DT)	98.99%
2016	Ashok Kumar Dwivedi	Evaluate the performance of different machine learning	Naïve Bayes	83%

		techniques for prediction of heart disease using tenfold cross-validation	Classification Tree	77%
			KNN	80%
			Logistic Regression	85%
			SVM	82%
			ANN	84%
2017	EmranaKabirHashi	An expert clinical decision support system to predict disease using classification techniques.	C4.5	90.43%
			KNN	76.96%
2017	HuseyinPolat et al.	Diagnosis of Chronic Kidney Diseasebased on SVM by feature selection methods	SVM	98.5% highest accuracy achieved by FilterSubsetEval with Best First
2017	MeghaShahi et al.	Heart disease prediction system using data mining techniques.	SVM, Naïve Bayes, Association rule, KNN, ANN and Decision Tree	The paper explains that in some paper SVM effective and efficient accuracy about 85% as compared to other data mining algorithms.
2017	Syed Muhammad Saqlain Shah et al.	Analysis of Heart Disease Diagnosis based on feature extraction using K- Fold cross-validation	SVM	91.30% is the highest accuracy obtained
2017	Curtis Langlotz et al.	Bone Tumor Diagnosis Using a Naïve Bayesian Model of Demographic and Radiographic Features	Naïve Bayesian	Using Naïve Bayes Differential accuracy 80% of accuracy is achieved
2017	Azam et al.	Automatic diagnosis of heart disease using K-Fold cross-validation method	Optimized SVM	99.20%

III. PROPOSED ME1THODOLOGY& RESULTS

The block drawing for organization of heart disease databank is shown in figure 1.

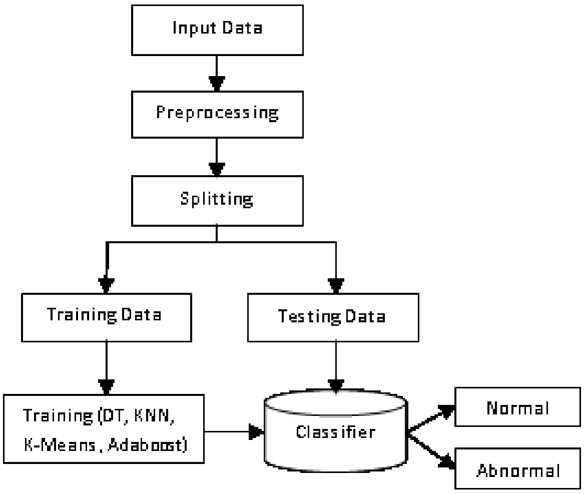

Fig.1 Proposed Data classification system

A. Preprocessing

The database contains NaN values. The NaN values cannot process by the programming hence these values need to convert into numerical values. In this approach mean of the column is calculated and NaN values are replaced by the mean.

B. Splitting

The whole database is split into training and testing database. The 80% data is taken for training while remaining 20% data is used for testing.

C. Classification

The training data is trained by using four different machine learning algorithms i.e. Decision Tree, KNN, K-mean clustering and Adaboost. Each algorithm is explained in detail.

1. Decision Tree

There are dissimilar kinds of decision trees. The only difference is in scientific ideal that they use to first-rate the class of feature through rule mining. A gain ratio decision tree is very common and fruitful category. It is the association amongst information gain and classified information.

In entropy system, the characteristic that reduces entropy and exploits information gain is nominated as tree root. For selecting tree root, it is first essential to estimate information gain of all attributes. Later, the attribute that exploits information gain will be nominated.

$$E = -\sum_{i=1}^{k} P_i log_2 P_i$$

Here k is count of response variable modules, piistheratio of the number of the ith class procedures to a total count of models.

2. *KNN*

This is one of the simplest and fundamental methods of classification where the user does have a little knowledge or no understanding of the dissemination of the data. While carrying out Discriminant examination when some dependable parametric controls of probability densities are not known or found challenging to understand this classification method was developed to perform such calculations.

The exact location of the K-nearest neighbor should be decided with the help of the training dataset. To find how much close each fellow of the training dataset is from the target how row that is to be examined, we make use of Euclidean distance. Discovery of the k-nearest neighbors and allocating the group to the row that is being inspected. Now repeat the technique for the rows outstanding in the target set. We can also select the maximum value of K in this software after that the software automatically builds a parallel model on the values of k upto the maximum specifies value.

The first phase by means of K-nearest Neighbor classification technique with the support of WEKA tool is to decide the training dataset and then the input and output variables must derive in. Standardizing the data is the second step it guarantees that the distance degree allocates identical weight to each variable is the second phase in this course. The best score achieved of k between 1 and the given value is chosen that helps building parallel models on all values of k up to the extreme identified value for which k=9 was selected and scoring is done using the finest models from the available ones. Finally the data needed for classification is entered.

3. *K-mean clustering*

It is an unsupervised learning which is used when class label is not known or you have unlabeled data. The main focus of this algorithm is finding the groups in the data with that number of groups that represent the variable K.

This algorithm iteratively allocating the k groups to the point. Data points here are clustered based on feature of similarity. The consequences of the *K*-means clustering algorithm are:

1) We can use centroid of the *K* clusters, to tag new data
2) The training data are tagged (A single data point is allocated to a single cluster)

Clustering defines groups beforehand observing at the obtainable data, and also allows us to diagnose and examine the groups that have been designed naturally. Each centroid of the accessible clusters is a group of feature ideals that defines the subsequent groups. By studying the centroid eye, weights can easily be used to qualitatively understand that the cluster fits to which group.

4. *Adaboost*

It is a fine technique that is used to increase the performance of decision tree on binary classification problems. AdaBoost was previously known as AdaBoost.M1. Currently it is also discussed to as discrete AdaBoost as it is used mainly for classification relatively than regression.

We can increase the presentation of every machine learning algorithm using Adaboost. It is finest used when the beginners are weak. These models gain the accuracy level just above the random chance on a given classification problem.

The common algorithm that is used with AdaBoost is decision tree but with one level. As these trees are tiny and can contain exactly one decision for classification, they are mostly called as decision stumps.

Each occurrence that is available in the training dataset should be weighted. The original weights are set to:

$$\boldsymbol{Weight}\ (X_i) = \frac{1}{n}$$

Where x_i is the ith training occurrence and n is the count of training occurrences.

IV. CONCLUSION

In the above paper we have studied various classification algorithms that can be used for classification of heart disease databases also we have seen different techniques that can be used for classification and the accuracy obtained by them. This investigation tells us about dissimilar technologies that are used in dissimilar papers with dissimilar count of attributes with different accuracies depending on the tools designed for execution.

The accurateness of the structure can be further upgraded by creating various combinations of data mining techniques and by parameter tuning also.

REFERENCES

1. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, “Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review” IJCA 2016.
2.]K.Sudhakar, Dr. M. Manimekalai “Study of Heart Disease Prediction using Data Mining”, IJARCSSE 2016.
3. NagannaChetty, Kunwar Singh Vaisla, NagammaPatil, “An Improved Method for Disease Prediction using Fuzzy Approach”, ACCE 2015.
4. VikasChaurasia, Saurabh Pal, “Early Prediction of Heart disease using Data mining Techniques”, Caribbean journal of Science and Technology,2013
5. ShusakuTsumoto,” Problems with Mining Medical Data”, 0-7695- 0792-1 I00@ 2000 IEEE.
6. Y. Alp Aslandoganet. al.,” Evidence Combination in Medical Data Mining”, Proceedings of the international conference on Information Technology: Coding and Computing (ITCC’04) 0-7695-2108-8/04©2004 IEEE.
7. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

8. Franck Le Duff, CristianMunteanu, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, page no. 1256-1259, 2004.
9. Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel,Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", Proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16,page no. 305-310, 2006.
10. Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.
11. Hongyu Lee, Ki Yong Noh, Keun Ho Ryu, "MiningBiosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, May 2007, page no. 56-66.
12. Niti Guru, Anil Dahiya, NavinRajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1, January - June 2007.
13. Hai Wang et. al.,"Medical Knowledge Acquisition through Data Mining", Proceedings of 2008 IEEEInternational Symposium on IT in Medicine and Education 978-1-4244-2511-2/08©2008 Crown.
14. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (IJCSNS), Vol.8 No.8, August 2008.
15. LathaParthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3,Page No. 3, 2008.
16. Chaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 888)Volume 47No.10, June 2012.
17. S. Vijiyarani et. al., "An Efficient Classification Tree Technique for Heart Disease Prediction",International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), 2013 (pp 6-9).
18. Harsh Vazirani et. al.," Use of Modular Neural Network for Heart Disease", Special Issue of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010 (pp 88-93).