Heart Diseases Forecast Using Data Mining Techniques and Tools

Karan Khayani, Sanchit Arora, I. Mala Serene

Abstract*: Data Mining have always been a field and combination of both computer science and statistical knowledge. From the beginning it is used to ascertain designs, patterns and arrangements which are formed in the information pool. The motive of the data mining development is to produce useful information from the pool of raw data and convert it into useful information which can be used for future arrangements. The tools which are used in data mining are helpful in predicting the future trends and predictions across the market, which also help in decision making and building the knowledge to make decisions. The "Healthcare Industry" is generally information rich. It has been collecting data to improve the continuing problems and help to identify the solutions for that problems. Data mining techniques can be used to predict heart conditions from the voluminous and complex data which are kept by the hospitals for decision making which are difficult to analyze by outmoded methods. Unfortunately, outmoded methods are less accurate in discovering hidden information from effective decision making. Data mining helps in altering the huge amount of data into knowledge driven which takes, as compared to others, less time and effort for the prediction and with greater accuracy. Our effort is to apply different data mining techniques that are used to solve the problem of biased forecasts and decision making and help in calculating the results with more accuracy.

Keywords: Heart disease, Data mining, Naïve Bayes, K-Nearest Neighbor, Decision Tree

I. INTRODUCTION

Data mining -a structural and procedural well-defined process which allows to extract datasets from a huge database and identify patterns and relationships between them [2]. Data mining is used for extracting or "mining" knowledge from the large pool of data. Outmoded database queries give direct result in the form of subset only which do not provide transparency whereas "Data Mining" gives output in the form of datasets which differs from the outmoded access [8]. Many hospitals use different systems but can only answer simple queries which gives normal subsets. It can't answer complex questions like possibilities although it can only be answered by doctor's experience and knowledge [5].

Data mining helps to return the used data driven information and remove unwanted errors and decrease the unwanted results and increases accuracy [2]. Data mining allows to extract knowledge based information from past records and database and helps in future predictions and market examination. It helps in identifying market trends which may fluctuate in future. The major use of data mining is in hospitals where large amount of data is used to predict the disease of the patients. In the same way, heart diseases are also being predicted by the use of data mining. The main motive of this paper is to predict heart diseases with suitable number of attributes like cholesterol, age, diabetes level etc. and with different classification techniques such as KNN, Naïve Byes and Decision Tree. Applying data mining techniques in heart disease will provide reliable and accurate information which is more sensitive and of higher degrees.There are several domains in medical wherever the maximum accuracy is desired. Cardiovascular diseases are the single largest reason behind death in developed countries and one in all the most contributors to disease burden in developing countries [1]. Unfortunately, the data available to the healthcare industries is not in the proper format which makes the extraction of information and hidden patterns difficult. However cardiovascular diseases or CVDs are on the hike in contribution to the death rate as compared to any other diseases. According to a World Health Organization survey an estimation was made that around 17.9 million people died from CVDs in 2016, surprisingly increasing, representing around 31% of all the total global deaths [7]. Out of these deaths, 85% are due to heart attack and stroke. There are several reasons like improper or lack in flow of blood to brain and heart which cause CVDs and others include high tension, smoking, fatness, stress etc. These attributes result in blocking of arteries of the heart which cause lack in blood flow to the main organs of the body [1].

II. LITERATURE SURVEY

Heart disease is a vast and major issue discussed everywhere and thus it can be predicted by taking some common attributes of a patient. Study can be done through many different tools and techniques which show several level of work done. Various data mining methods are being used for treatment and diagnosis, and have achieved different but reliable accuracy levels. The data mining classification techniques such as K-NN, Naïve Bayes and Decision Tress are used to make the predictions and are tested on heart disease dataset.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author

Karan Khayani, School of Information Technology and Engineering, VIT Deemed to be University, Vellore, India.
Sanchit Arora, School of Information Technology and Engineering, VIT Deemed to be University, Vellore, India
Mala Serene I*, School of Information Technology and Engineering, VIT Deemed to be University, Vellore, India.

With the increase in technology and its applications day by day, data mining prompts a very major and key role in early detection of diseases. Data mining methods can be implemented for predicting the results of the domain. Hence prediction plays a very important role [4].

The methodology and implementation of Naïve Bayes and Decision Tree technique used for prediction of heart diseases is discussed in [1]. The results show the short but accurate outcomes by applying classification which is used for the newly added patients compared to the already added ones. A model based on Combination of Naïve Bayes Classifier and K-Nearest Neighbor is proposed in [3, 4, 11]. Adding two more attributes helps in defining the clear result. The functioning of heart disease prediction using actual real time data from health care organizations and agencies which can be built using big data is proposed in [8]. Their model which can be expanded shows higher accuracy extracted from the hidden knowledge. A model based on K-nearest clustering which is user for large dataset is shown in [9]. Their paper shows the real time investigation for newly added patients with higher accuracy. Prediction of risk factors and symptoms with statistical data is done by World Health Organization (WHO) report in 2018 [7].

III. PROPOSED PREDICTION SYSTEM

Today many healthcare industries use data mining techniques to extract the hidden information from the data pool to come to a right decision. The diagnosis is based on the prediction of the hidden data which has to be accurate. The main objective is to build a legit system which is used to do diagnosis accurately [5]. To develop such system, medical terms like obesity, sex, sugar level and blood pressure and such 13 other attributes are used. Classification techniques like Decision Tree and Naïve Bayes and K-NN are used.

IV. DATA SOURCE

The dataset on which data mining techniques have been implemented is available publicly for direct use. This data set contains 303 records with 13 attributes. These attributes can be seen almost in all papers for prediction of heart disease. A target attribute is added to target the outcome and compare with the available training data set.

V. TECHNIQUES FOR PREDICTION(S)

The three major and popular data mining techniques are used which are Naïve Bayes, Decision Tree and K-NN.

A. K-Nearest Neighbors (K-NN): K-NN, which is also known as "lazy algorithm" is the simplest classification algorithm. It is very much popular technique because of its accuracy level. K-NN checks the upcoming cases with the stored ones based on similarity symptoms [5]. It has been used in classifying that in which group the new data exists.

Table 1: Available attributes [5]

1	age	Age in years	Continuous
2	sex	Male or female	1=male, 0=female
3	cp	Chest pain Type	1=typical type1 2= typical type agina 3=non-agina pain 4=asymptomatic
4	thestbps	Resting blood pressure	Continuous value in mm hg
5	chol	Serum Cholesterol	Continuous value in mm /dl
6	Restecg	Resting electrographic results	0=normal1=having_ST_T wave Abnormal2=left ventricular hypertrophy
7	fbs	Fasting Blood sugar	1>=120mg/dl 0<=120mg/dl
8	thalach	Maximum heart rate received	Continuous value
9	exang	Exercise induced aging	0=no 1=yes
10	oldpeak	ST depression Induced by exercise relative to rest	Continuous value
11	solpe	Slope of the peak exercise ST segment	1=unsloping 2=flat 3=downsloping
12	ca	Number of major vessels colored by floupoy	0-3 value
13	thal	Defect type	3=normal 6=fixed 7=reversible defect
14	target	target	1=yes 0=no

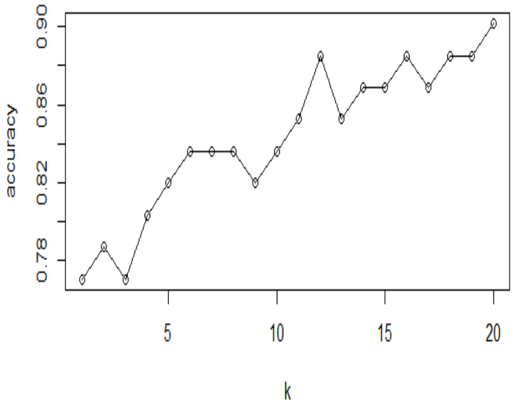

Fig. 1 Accuracies at different values of K

Whenever a prediction is required, it searches for the similar K value in the entire dataset and return the most significant outcome. It says that if your are similar to a group then your are one of them. In figure 1, accuracies at different K values are calculated and depicted. It is clearly shown that as the K value is increasing, the accuracy is also increasing.

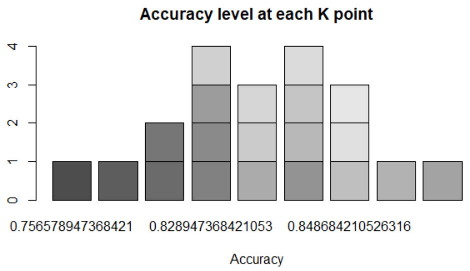

Fig. 2 Accuracies at different values of K using Bar diagram

Though K-NN is classic clustering technique nevertheless widely used for data mining as it gives more accuracy and take less time to cluster the data as compared to other techniques.

B. Naïve Bayes Classification: Naïve Bayes classifier which is used for calculating estimations is based on a mathematical concept “Bayes theorem”. It uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes [5]. Given two events A and B, P(A) is prior probability and P(A|B) is posterior probability, then according to Bayes theorem P(A|B) is expressed as:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)}$$

C.Decision Tree: Decision tree approach is more powerful for classification problems. The main tree is divided into sub trees indicating each branch as an outcome or a result and each leaf node depicts a label. The tree is applied to the dataset and the results are made to the least [5]. J48 is applied in this paper which is the simplest and requires no as such domain knowledge and also it reduces errors. Results from various papers shows that decision tree is more powerful but it takes more time to build a tree.

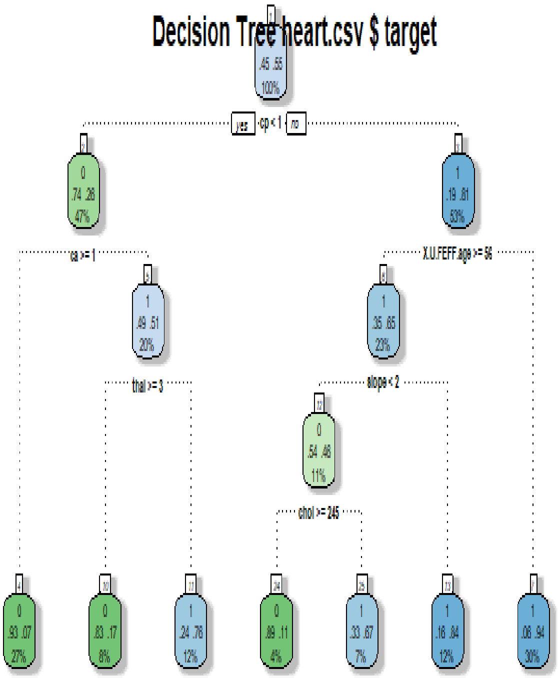

Fig. 3. Decision tree using J48 algorithm

VI. METHODOLOGY

Heart disease forecast using data mining can be appreciated in figure 4. Raw dataset is collected and used and then a pre-processing technique is applied to provide the correct dataset. Any classifier technique is being used with the dataset and is trained with the data. In this paper, techniques like Naïve Bayes, Decision Tree and K-NN is used [3]. The provided dataset is then split into two - train dataset and test dataset by a percentage criterion. In this paper the main dataset is divided into 80% with training data and remaining 20% with test data. By applying pre-knowledge on the test dataset and by applying techniques the outcome is predicted and is compared with the training dataset [5]. This model can be used in healthcare industries for accuracy concerns regarding heart diseases.

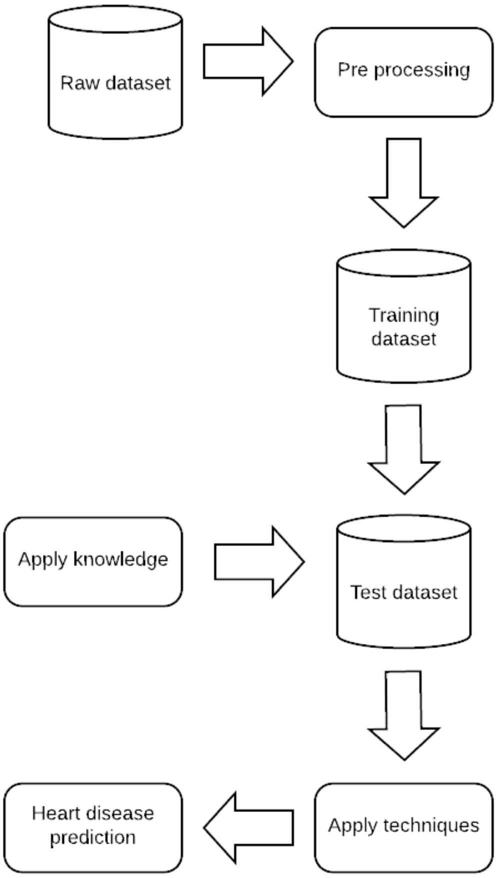

Fig. 4. Methodology for heart disease forecast

VII. RESULT AND DISCUSSION

The dataset being implemented consists of 303 records beyond which 80% records are trained for training dataset and remaining are trained for test dataset. The data mining tool R 1.2.5001 is used to build this system.

Initially the dataset contains some values like NULL, which are not to be calculated, which are removed with some calculable values. This is known as data preprocessing. Then classification techniques such as Naïve Bayes, Decision Tree and K-NN are used [3].

A confusion matrix is prepared to estimate the accurateness of the predictions.

Its show the total number of instances or the target variable allocated to each class [5]. In our trial two classes have been used.
Class A=1 (has heart disease)
Class B=0 (doesn't have heart disease)

Table 2: Confusion matrixes obtained yet with 13 attributes

Tab. 2.1 Confusion matrix of K-NN

	0	**1**
0	23	6
1	5	27

Tab. 2.2 Confusion matrix of Decision Tree

	0	**1**
0	21	4
1	8	30

Tab. 2.3 Confusion matrix of Naïve Bayes

	0	**1**
0	110	5
1	10	145

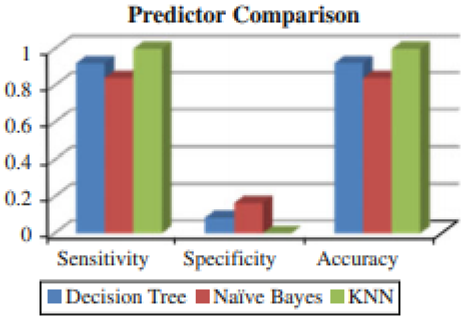

Fig. 5. Graphical representation for comparisons for each method [4]

Though the result can be predicted from any algorithm but the more accurate an algorithm could be the more it is usable [4]. From figure 4 it can be clearly predicted that the K-NN is more accurate and sensitive than others. Also it takes much lesser time to cluster. On the other hand, Naïve Bayes shows the maximum belonging and relatively uniquely quality. Clustering is the major topic to perform the techniques fair in every sense. Figure 4 show K-NN performs more rigid clustering hence it is a legit cluster method. As a result, K-NN is also useful in large datasets.

VIII. CONCLUSION

The objective of our paper is to execute data and forecast the outcome with more and better accuracy. Techniques like Naïve Bayes, Decision Tree and K-NN are used to give results with more accuracy do that is can lead to predict with more accuracy [4]. Although all three techniques are being compared together to check which is more accurate.

This system is built in such a way that it can be expanded by adding with more number of attributes. In some papers it is shown that Naïve Bayes gives more accuracy whereas in some paper it is written that K-NN gives more [4]. So, different techniques are shown in our paper to tell which is more accurate in predicting the outcomes. Though K-NN is famed for its accuracy and less time consumption whereas as Naïve Bayes is for more specificity. Although data mining techniques helps the healthcare industries in some aspect, the success of this system is useful to the patients suffering from heart diseases which has gathered more attention.

REFERENCES

1. M.Akhil jabbar, Priti Chandrab, B.L Deekshatuluc, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies(ICECIT).2012.
2. Vikas Chaurasia, Saurabh Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology.,vol. 2, no. 4, pp. 56-66 ,2013.
3. Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International Journal of Computer Science and Network Security., vol.8, no.8, pp. 343-350,2008.
4. Joshi S., Nair M.K. (2015) Prediction of Heart Disease Using Classification Based Data Mining Techniques. In: Jain L., Behera H., Mandal J., Mohapatra D. (eds) Computational Intelligence in Data Mining,vol.2. pp 503-511, 2015
5. Chaitrali S. Dangare , Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications.,vol.47,no.10, 2012.
6. Beant Kaur, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", International Journal on Recent and Innovation Trends in Computing and Communication,vol.2,no.10,pp.3003-3008, 2014.
7. World health organization report, 2018.Available: https://www.who.int
8. Shinde S., Amrit Priyadarshi, "Diagnosis of Heart Disease Using Data Mining", International Journal of Science and Research (IJSR), vol.4.no.5, pp.2301 – 2303, 2015.

AUTHORS PROFILE

Karan Khayani, undergoing Master of Computer Applications, School of Information Technology at Vellore Institute of Technology, Vellore, is in his first year of graduation.

Sanchit Arora, undergoing Master of Computer Applications, School of Information Technology at Vellore Institute of Technology, Vellore, is in his first year of graduation.

I. Mala Serene working as Associate Professor & HOD, Smart Computing in School of Information Technology & Engineering (SITE) at VIT University, Vellore. She received her B.E degree in Electronics and Communication Engineering in Karunya Institute of Technology and M.Tech degree in Computer Science and Engineering from VIT University during 1992 and 2002 respectively. She completed her Ph.D at VIT University, Vellore with good number of publications indexed by Scopus. She is having around 22+ years of teaching experience and 4 years on Industrial Experience in Instrumentation field. She is a life time member in Computer Society of India (CSI), ISTE and annual member of Indian Science Congress Association (ISCA). Her areas of interests include MEMS and Data Mining.

Retrieval Number: B6594129219/2019©BEIESP
DOI: 10.35940/ijitee.B6594.129219
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication