A STUDY ON PREDICTION OF HEART DISEASE USING DATA MINING TECHNIQUES

Kanika Khera , Dr. Neelam Duhan
PG Student , Assistant Professor
Information Technology
YMCAUST, Faridabad, Haryana, India

Abstract:

Data mining is the evolution in information technology field to extract large amount of data and apply it efficiently to solve real life problems with the help of different tools and techniques. To being with, there is a large amount of raw data in the pre-processing stage. Data mining techniques are then applied on that data to discover patterns and logical links among them which in turn helps in extracting useful information and knowledge from the it. For the processing of medical data, there are various data mining tools and techniques which help professionals in making decisions based on the logical inferences suggested by these tools.

This survey explored various papers and research-work for prediction of heart disease with the help of machine learning algorithms. Heart disease can be related to a large number of medical conditions. By the use data mining technique, it will be more easier to predict and observe disease with higher accuracy. In data mining, there are some well-known techniques which can be applied here. Those techniques are classification, clustering, prediction, sequential patterns & association rules.This study predominantly examines various data mining classification algorithms used in cardiac disease predictions.

Keywords: Data mining, Heart Disease prediction, Data mining techniques, Accuracy.

1. INTRODUCTION

Data mining is used to analyze large amount data and derive useful knowledge from it. This knowledge then can be applied in various real life applications such as in healthcare industry. The biggest challenge for any technology when healthcare industry is involved is the level of accuracy it can provide. Even a slight compromise is unacceptable in most of the cases as the incorrect results can lead to wrong treatment of patients and in the worst case it can be fatal as well.

The data mining techniques can help in diagnosing diseases very early and in cost effective manner. The heart attack is one the major reasons of pre-mature deaths across the globe,especially in developing countries. Traditional Medical diagnosis find it difficult to predict cardiac health with accuracy and efficiency but it is very important to have that knowledge so that patients can be treated on time and in a best possible manner .

The dataset should be organized according to hospital management system.

Some of the pre-dominant heart ailments are cardiovascular diseases, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease which is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure(6). Heart disease kills one person in every 34 seconds in the United States(8).

Popular data mining techniques which are used in the diagnosis of heart disease are the logistic regression, Naïve Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing distinct levels of accuracies.

This paper presents the data mining techniques in detail. These techniques will help both patients and doctors by saving time, accuracy and cost reduction.

2. OVERVIEW OF HEART DISEASE

Heart plays a very important role in the overall functioning of the body and even a slight deterioration in its health reduces the overall quality of human life.If unattended for a long time,heart ailments becomes fatal.According to a WHO report , around 15 million death occurs every year because of heart diseases.

The main factors which are responsible for the coronary disorder are classified as controllable risk factors and uncontrollable risk factors. Some of the controllable risk factors include:

- Diabetes
- Smoking
- Obesity or excess weight
- Cholesterol
- High blood pressure
- Lack of exercise
- Physical activity Uncontrollable risk factors include:

- Genetic
- Age
- Sex
- Previous medical disorders.

3. DATA MINING TECHNIQUES USED

There are different data mining techniques for classification [3].Performance analysis on different classification algorithms such as Decision tree, Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Neural Networks (NN) are carried out.

limitations in terms of the type of data it handles, accuracy, and ease of understanding, reliability and generalization ability. Various data mining techniques such as Naïve Bayes, Support vector machine and decision tree are used [4]. Another method suggest to use data mining techniques such as Genetic Algorithm, Support Vector Machine (SVM), association rules, rough set theory and Neural Networks. Out of the above techniques Decision Tree and SVM is most effective for the heart disease. For future work, more Accuracy can be increased by increasing the attributes by using different data mining techniques [5]. This [6] paper discusses and presents the experiment that was executed with Naïve Bayes technique in order to build predictive model as an artificial diagnose for heart disease based on data set which contains set of parameters that were measured for individuals previously.

Accuracy of the naïve bays model achieved ratio (100%). The three data mining techniques are used in this paper such as CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) and also evaluated the performance of the three classifiers. Different classifiers are studied and the research is conducted to find the best classifier for calculating the patients Diagnosis. The best algorithm is CART which gives more accuracy [7]. This paper uses K-means clustering technique which is applied to find out clusters in data which are further used to remove hidden forms related to heart patients. In future work, they have planned to implement an expert system that would predict the probability of patient being Critical or at risk state using logistic regression algorithm and these extracted patterns [8].

Naïve Bayes : -

Naïve Bayes is one of the popular data mining methods. This algorithm is used to create predictive models. A more descriptive term which can be used is "independent feature model". In Naïve Bayes two classes are independent of each other. For example, a vehicle may be considered to be a car if it has 4 tiers , steering, and covered. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this vehicle to be a car. The main advantage of Naïve Bayes it requires only small amount of training data for classification. Because, independent variables only need to find variance of variable for each class instead of covariance matrix. In Fig (1) naïve bayes described in mathematical form.

$$\underset{\text{Posterior Probability}}{P(c \mid x)} = \frac{\overset{\text{Likelihood}}{P(x \mid c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

ORIGINAL BELIEF
+
OBSERVATION
NEW BELIEF

(1)

SVM:
A Support Vector Machine model is used in classification and regression for analyzing and discover patterns. SVM performance is effective on huge amount of dataset.Support Vectors are simply the co-ordinates of individual observation. For instance, (45,150) is a support vector which corresponds to a female. Support Vector Machine is a frontier which best segregates the Male from the Females. In this case, the two classes are well separated from each other, hence it is easier to find a SVM.

Then the probability (p) of each case is calculated using odds ratio, P/(1-P) = eY – (2)
From this p-value will be calculated. This P-value will find the probability(p) of having heart disease or not.

Logistic Regression:-
Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

Regression is formulating a functional relationship between a set of independent or Explanatory variables (Xs) with a dependent or Response variable(Y).
Y=f(X1,X2,…,Xn)
The probability of logistic regression is between 0 and 1, that is, S-shaped. Logistic regression models logit of outcome.

ROC Curve- AUC

ROC is the overall measure of model performance.

Decision Trees:-
The decision tree approach is efficient to use for classification problems. This is solved in two steps

1. building a tree ,

2 .applying the tree to the dataset.

The decision tree algorithms are CART, ID3, C4.5, CHAID, and J48..This technique gives maximum accuracy on training data.

4. LITERATURE SURVEY

Dataset Used

In this heart disease prediction,, datasets were collected from UCI machine learning repository.. The dataset for heart disease consists of 14 attributes with 270 instances. Values of attribute can be continuous and categorical. On that dataset classification algorithms were applied to determine the accuracy and the performance for disease prediction.

Table1: Dataset Description

S.No	DATASET	ATTRIBUTES	INSTANCES	TYPE
1.	Heart Disease	14	303	Numeric & Nominal

The main objective of this study is to predict heart disease using risk factors like Age, Sex, Chest pain type, Resting blood sugar, Cholesterol, Resting Electrographic results, Fasting blood sugar, Thalach, Exang, Oldpeak, Slope, Number of major vessels colored by Flourosopy, Thal, Height and Weight. In proposed research pre-processing techniques includes noise removal, discarding records with missing data, filling default values if applicable and classification of attributes for decision making at different levels. In this survey, there are classification techniques and their accuracy. In fig. 1, flow of work for prediction of heart disease using SVM and Naïve Bayes In table 1 , data mining classification techniques like Support Vector Machine (SVM) and Naïve Bayes are used.[9]

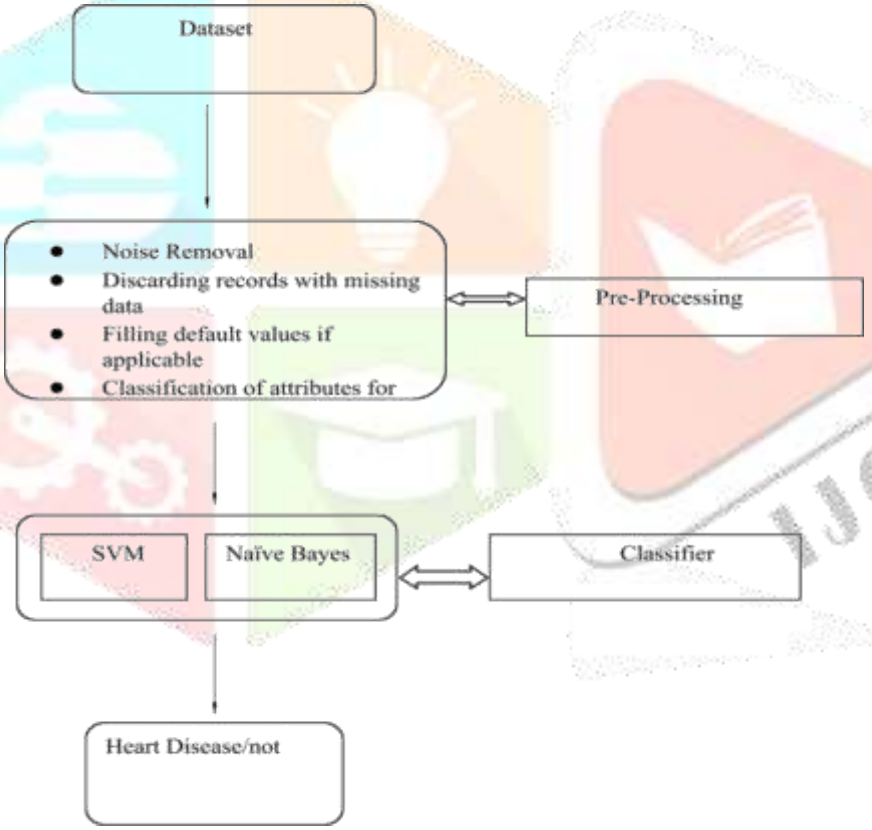
Dataset
Noise Removal
Discarding records with missing data
Filling default values if applicable
Classification of attributes for
Pre-Processing
SVM
Naïve Bayes
Classifier
Heart Disease/not

Fig1: prediction of cardiovascular Disease

Table 2: Results of classification algorithm

Training & Testing Ratio	SVM			Naïve Bayes		
	Accuracy	Specifity	Sensivity	Accuracy	Specifity	Sensivity
50:50	57%	43%	38%	61%	50%	50%
70:30	53%	85%	25%	49%	65%	23%
75:25	55%	80%	25%	40%	84%	19%
80:20	57%	87%	35%	52%	86%	28%

Classification using C5.0

a) Rule 1, testing phase: 50% of instance were selected from both the dataset. It can be used to attain the classification rule sets.
b) Rule 2, In Testing Phase & Performance Analysis: To determine the classifier's accuracy, the generated classification rules were applied to the testing data. And then the actual and predicted values will be generated from which confusion matrix is fetch. [10]

Table 3 : Confusion Matrix Representation

	A	B
A	True Positive	False Negative
B	False Positive	True Negative

Sensitivity = TP / TP + FN
Specificity = TN / TN + FP
Precision = TP / TP + FP
True-Positive Rate = TP / TP + FN
False-Positive Rate = FP / FP + TN
True-Negative Rate = TN / TN + FP
False-Negative Rate = FN / FN + TP

Confusion matrix can be used to analyze the accuracy, which is the performance measure of a classifier.

Table 4 : To measure Accuracy and Error rate using C5.0

Dataset	Accuracy	Build	No. of	Error Rate

	Training Set (%)	Testing Set(%)	time	Attributes used	Training set (%)	Testing Set(%)
Heart Disease dataset	92.248%	76.596	<1min	8 out of 13	0.07	0.23

SVM Algorithm & CNN Algorithm:
The training and testing dataset accuracy of CNN is 78.55% and 85% respectively. The training and testing dataset accuracy is 75% and 84% respectively
In neural network when desired accuracy achieved the system will stop immediately.

The accuracy of CNN is increased by almost 3% .

It can be noticed from the figure accuracy of CNN is more as compared to SVM.[12]

Table 5: performance measure for CNN & SVM

Measure	**CNN**	**SVM**
Accuracy	0.85	0.82
Sensitivity	0.83	0.87
Specificity	0.87	0.775

Artificial Neural Network:
In machine learning, artificial neural networks (ANNs) is the learning model encouraged by biological neural networks (i.e. nervous systems of humans brain) and are used to predict or analyze functions that can depend on a large number of inputs. Artificial neural networks are normally offered as systems of interconnected "neurons" which trade messages between each other. The connections have numeric weights that can be tuned based on experience, making training, testing adaptive to inputs and capable of learning. Naïve Bayes is a classic Bayesian theorem for classification with strong independence assumption. It gives best results when the input is multi-dimensional. It gives a probability that a particular class of a tuple can be predicted using values entered. This method was proposed by Thomas Bayes which is given by following formulae: $p(h|e)=[p(e|h)*p(h)]/p(e)$ The Hypothesis or eventh is calculated based on the evidences (e) that are observed. Model proposed by naïve bayes uses conditional probability, the test case or a tuple which is to be classified is represented by a vector x(x1,x2,…xn). The different values of x represents n independent features, for k possible classes or outcomes, p(ck| x1,x2,…xn). If number of features n increases or n take large number of values it classification becomes difficult, therefore the above formula is remodel to, $P(ck|x) = P(ck)*P(x|ck)/P(x)$ Where, P(ck|x) is posterior probability P(ck) is prior probability P(x|ck) is likelihood And P(x) is evidence.
In comparison , both from ANN and Naïve Bayes the accuracy of Naïve Bayes is higher.[11]

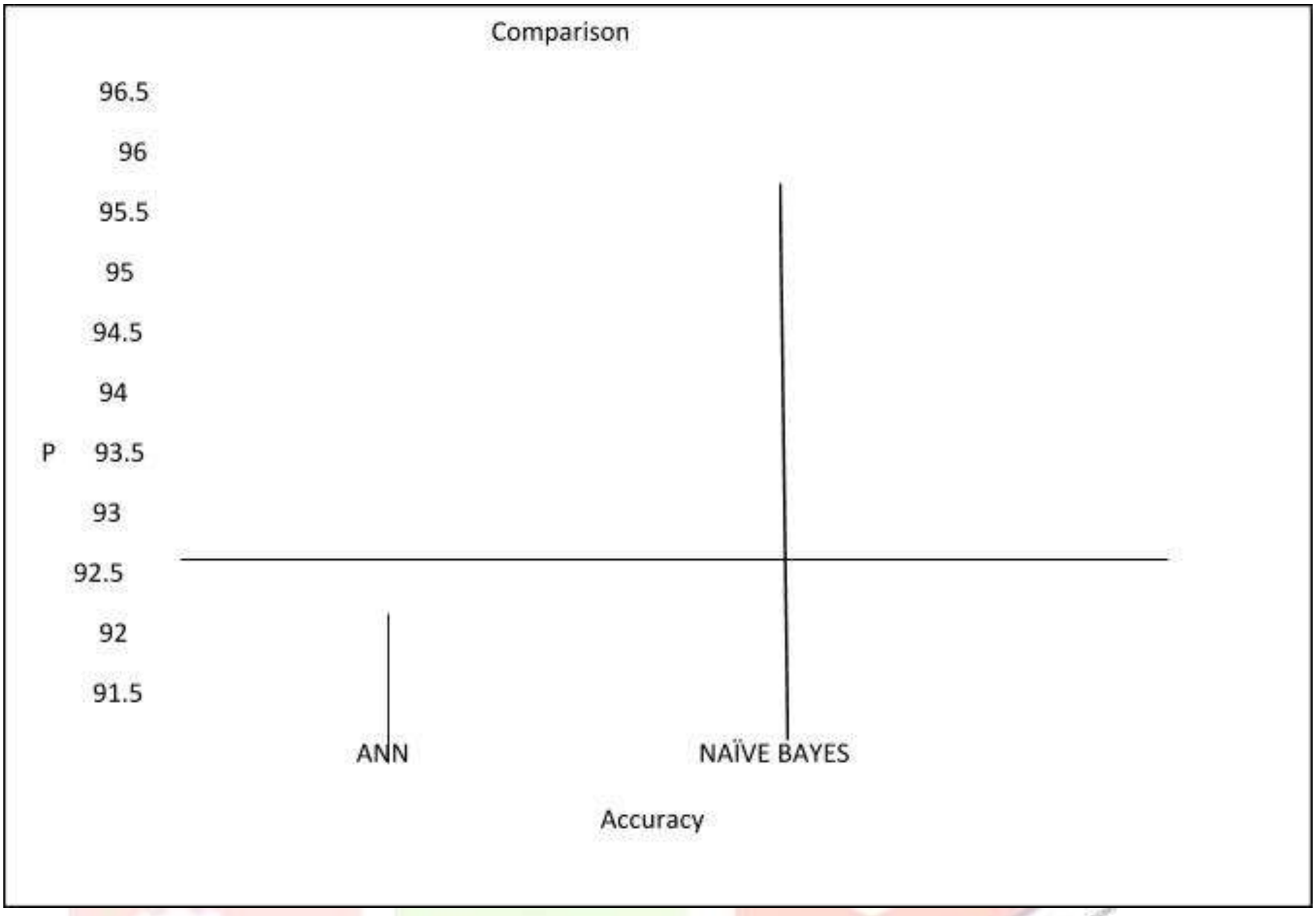

Fig 2: comparison of algorithms

5. CONCLUSION AND FUTURE WORK

The objective of this study work is to provide a different data mining techniques which is helpful for automated heart disease prediction systems. With the help of various techniques and data mining classifiers heart disease prediction become efficient and effective. The analysis shows that different techniques are used in various papers with taking different number of attributes. So, different techniques are used to shown the different accuracy. This can be enhanced on complex dataset and different machine algorithms used to increase its accuracy.

Acknowledgment

The success and final outcome of this study required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my study. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I owe my deep gratitude to my supervisor Dr. Neelam Duhan who took keen interest in my work and guided me all along, till the completion of our study by providing all the necessary information and made it possible to complete it on time.

I would not forget to remember my family for their encouragement and more over for their timely support and guidance till the completion of my work.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of YMCA which helped us in successfully completing of work.

(Kanika Khera)

REFERENCES

[1]Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review(International Journal of Computer Applications (0975 – 8887) Volume 136
– No.2, February 2016)

[2]Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques.

[3]T.John Peter, K. Somasundaram," An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", IEEE, International conference on Advances in engineering, science and management,pp.514-518, 2012.

[4]Eman AbuKhousa, Piers Campbell," Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", IEEE, International Conference on Innov

[5]Aqueel Ahmed, Shaikh Abdul Hannan," Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012.

[6]Lamia AbedNoor Muhammed," Using Data Mining technique to diagnosis heart disease", IEEE, International conference on statistics in science, Buiseness and Engineering, pp.1-3, 2012.

[7]Vikas Chaurasia, Saurabh Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech, 2013, Vol.1, 208-217.
[8]Mamuna Fatima, Iqra Basharat, Dr. Shoab Ahmed Khan, Ali Raza Anjum,, "Biomedical (Cardiac) Data Mining: Extraction of significant patterns for predicting heart condition", IEEE conference on Computational Intelligence in bioinformatics and computational biology, pp.1-7, 2014.

[9]International Journal of Computer Applications (0975 – 8887) Volume 156 – No 2, December 2016 9 Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification

[10]International Journal of Computer Applications (0975 – 8887) Volume 155 – No 8, December 2016 20 Analysis of Classification Techniques for Efficient Disease Prediction.

[11] JSRD - International Journal for Scientific Research & Development| Vol. 4, Issue 03, 2016 | ISSN (online): 2321-0613 All rights reserved by www.ijsrd.com173 A Heart Disease Prediction System using Artificial Neural Network and Naive Bayes Ashish Bhatia1 Mohit Matwani2 Pawan Karira3 Rahul Sidhwani4 Asha Bharambe5 1,2,3,4Student 5Professor 1,2,3,4,5Department of Computer Engineering 1,2,3,4,5V.E.S. Institute of Technology, Chembur, Mumbai

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