International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A NOVEL TECHNIQUE TO PREDICT HEART DISEASES USING DATA MINING

[1]**Divya Kundra**
Shaheed Udham Singh College of
Engineering & Technology, Tangori, Distt Mohali.
divyanujdhir@gmail.com
[2]**Er. Navpreet Kaur**
Shaheed Udham Singh College of
Engineering & Technology, Tangori, Distt Mohali.
er.nrupal@gmail.com

Abstract: *Data mining is the process of analyzing large sets of data and then extracting the meaning of the data. It helps in predicting future trends and patterns, allowing business in decision making. Data mining applications can answer business questions that take much time to resolve traditionally. Large amount of data which is generated for the prediction of heart disease is analyzed traditionally and is too complicated and voluminous to be processed. Data mining provides the techniques and methods for the transformation of data into useful information for decision making. These techniques make the process fast and it takes less time for the prediction system to predict the heart disease with more accuracy. In this paper we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. Result from using neural networks is 99.62% in one paper [6].By Applying data mining techniques to heart disease data which needs to be processed, we can get effective results and achieve reliable performance which will help in decision making in healthcare industry.*

Keywords: *Heart disease ,data mining, data mining techniques, data mining tools, data mining applications, methodology.*

I. INTRODUCTION

Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown patterns and trends in databases. This information is further used to build predictive models. The main objective of our paper is to learn the different data mining techniques which are used in the prediction of heart diseases using any data mining tool. Heart is the most vital part of the human body as life is dependent on efficient working of heart. A Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol. Heart diseases can be caused due to number of factors:

High blood pressure: when the heart pumps blood, the force of the blood pushes against the walls of the arteries causing pressure. If the pressure rises and stays high over the time it is called high blood pressure or hypertension which can harm the body in many ways i.e. Increasing the risk of heart stroke or developing heart failure, kidney failure etc.

High cholesterol: cholesterol is a waxy substance found in the fatty deposits in the blood vessels. Increase in the fatty deposits (high cholesterol) does not allow sufficient blood to flow in through the arteries causing heart attacks.

Unhealthy diet: eating too much fast food increases blood pressure and cholesterol level causing the risk of heart attacks.

Smoking: it damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks.

Lack of physical activity: lack of exercise increases cholesterol level in blood vessels which further increases the risk of heart attacks. Obese people are more likely to have high blood pressure, high cholesterol level and diabetes (increase in blood sugar level) which increases the risk of heart strokes in human body. Nowadays, data mining is gaining popularity in health care industry as this industry generates large amount of complex data about hospital resources, medicines, medical devices, patients, disease diagnosis etc. This complex data needs to be processed and analysed for knowledge extraction which will further help in decision making and is also cost effective.

World health organisation has estimated 17.5 million people died from cardio vascular diseases in 2012, representing 31 percent of all global deaths. Out of these, an estimated 7.4 million were due to coronary heart disease and 6.7 million were due to stroke. WHO estimated by 2030, almost 23.6 million people will die due to heart disease as written in [3].

Thus, a beneficial way to predict heart diseases in health care industry is an effective and efficient heart disease prediction system. This system will find human interpretable patterns and will determine trends in patient records to improve health care.

II. RELATED WORK

Over the years, numerous works have been done related to heart disease prediction system using different data mining algorithms by different authors. They tried to achieve efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. This paper aims at analyzing different data mining techniques that has been introduced in recent years for heart disease prediction system by different authors.

M.A.Nishara Banu and B.Gomathy et al.[7] used C4.5 algorithm, MAFIA and K means clustering in the year 2014 using 13 attributes in the dataset achieving 89 percent accuracy.

Aqueel Ahmed et al. [2] show the classification techniques in data mining and show the performance of classification among them. In this classification accuracy among these data mining has discussed. In this decision tree and SVM perform classification more accurately than the other methods and was able to achieve 91% accuracy.

Abhishek et al in the year 2013 used data mining tool Weka 3.6.4 in heart disease prediction system usingJ48 technique achieved 95.56% accuracy and using Naive Bayes achieved 92.42%. [1]

Ms.Ishtake et al. [5] developed a prediction system for heart diagnosis using decision tree, Neural Network and Naive Bayes techniques using 15 attributes in the year 2013.

Chitra R.et al. [4] developed the computer aided heart disease prediction system that helps the physician as a tool for heart disease diagnosis. From the analysis it is concluded that neural network with offline training is good for disease prediction in early stage an d good performance can be obtained by pre-processed and normalized dataset.

Nidhi Bhatla et al. [6] projected the study of different data mining techniques that can be employed in automated heart disease prediction systems. The analysis shows that Neural network with 15 attributes has shown the highest accuracy. On the other hand, Decision tree has also performed well with 99.62% accuracy by using 15 attributes.

Shadab et al. [9] used Naive Bayes technique in the year 2012 using 15 attributes in the dataset for the heart diagnosis in heart prediction system.

Rashedur et al in the year 2013 used Neural network technique using Weka data mining tool and achieved 79.19% and to compare various classification techniques, he used another technique fuzzy logic using TANGRA data mining tool and achieved 83.85% accuracy. [8]

III.MATERIALS AND METHODS

Data mining algorithms that are being used in the proposed work are K nearest neighbour, naive bayes, linear discriminant analysis and SVM i.e. Support Vector Machine algorithms.

K Nearest Neighbor Algorithm : In k Nearest Neighbor algorithm object is assigned to the class which is most common in its neighbors. This algorithm is mathematical computational algorithm and is used for binary classification i.e. 0 & 1. It works best where the data has exactly two output classes.The input consists of the k closest training examples in the feature space. In k-nn classification, the output is a class member. An object is classified by the majority of its neighbors , with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor**.**

Linear Discriminant Algorithm: Linear Discriminant Algorithm finds linear combination of features that separates two classes of objects. Linear discriminant Analysis is a generalization of Fischer’s linear discriminant, a method which is used in machine learning to find a linear combination of features that seperates two classes of objects or events. It attempts to model the difference between the two classes of data.

Naive Bayes Algorithm: Naive Bayes algorithm considers each of the feature to contribute independently to the probability that the person has a heart disease. Naive Bayes classifiers assumes that the value of a particular feature is independent of the value of any other feature. It is used for binary classification i.e. 0 & 1 since it is a mathematical computational algorithm. This algorithm is very stable as a small change in the data set does not make a big change in the model.

Support Vector Machine Algorithm: Support vector machine algorithm works best where the data has exactly two classes. It finds the best hyper plane that separates the data points of one class from the data points of another class. It is best supported for binary response variables i.e. 0 & 1. Hyper plane is the separation between the two critical points or members (also called support vectors).

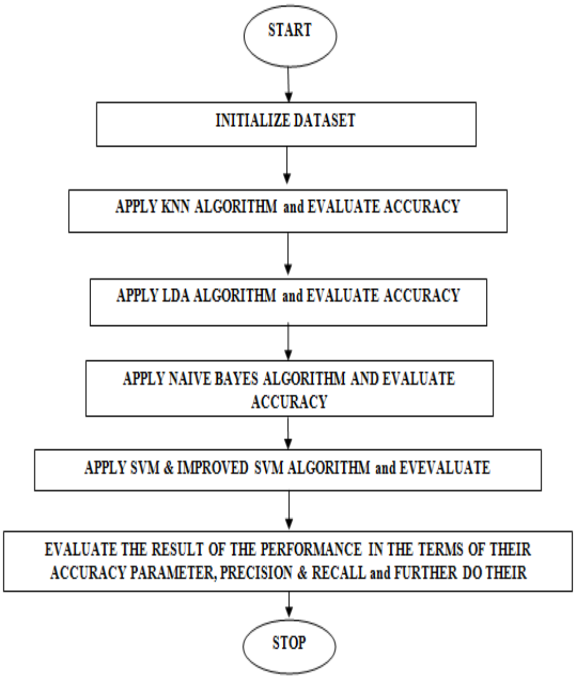
Improved Support Vector Machine: In machine learning, the (Gaussian) radial basis function kernel or RBF kernel is a popular kernel function used in various kernelized learning algorithms. We are improving support vector machine algorithm by applying this RBF kernel trick by approximations to RBF kernel.

Table 1- Heart disease attributes

Id.	Attribute
1.	Age
2.	Blood cholesterol
3.	Blood pressure
4.	Hereditary
5.	Smoking
6.	Alcohol intake
7.	Physical activity
8.	Diabetes
9.	Diet
10.	Obesity
11.	Stress
12.	Sex

Data Mining Tool

Matlab : Matlab is matrix laboratory. It is a multi paradigm numerical computing environment & fourth generation programming language. A proprietary programming language developed by Math Works, Matlab allows matrix manipulations, plotting of functions & data, implementation of algorithm, creation of user interfaces, and interfacing with programs written in other languages c, c++, java, Fortran & Python etc.


START
INITIALIZE DATASET
APPLY KNN ALGORITHM and EVALUATE ACCURACY
APPLY LDA ALGORITHM and EVALUATE ACCURACY
APPLY NAIVE BAYES ALGORITHM AND EVALUATE ACCURACY
APPLY SVM & IMPROVED SVM ALGORITHM and EVEVALUATE
EVALUATE THE RESULT OF THE PERFORMANCE IN THE TERMS OF THEIR ACCURACY PARAMETER, PRECISION & RECALL and FURTHER DO THEIR
STOP

Flow Diagram of proposed system

IV. METHODOLOGY

INITIALIZE THE DATASET: the dataset is mined, uploaded and transformed into the required matrix form with the help of data mining tool Matlab.

APPLY K NEAREST NEIGHBOR ALGORITHM AND EVALUATE ACCURACY: K nearest neighbour algorithm is applied to the dataset and accuracy is calculated. This algorithm assigns the object to the class which is most common in its neighbors.

APPLY LINEAR DISCRIMINANT ALGORITHM AND EVALUATE ACCURACY:
Linear discriminant algorithm is applied to the dataset and accuracy is calculated separately. this algorithm finds linear combination of features that separates two classes of objects.

APPLY NAIVE BAYES ALGORITHM AND EVALUATE ACCURACY:
This algorithm is applied to the dataset again separately and then the accuracy of this algorithm is calculated. Naive Bayes considers each of the features to contribute independently to the probability that the person has heart disease.

APPLY SUPPORT VECTOR MACHINE ALGORITHM AND EVALUATE ACCURACY:
Support vector machine algorithm is applied to the dataset and accuracy is calculated. This algorithm finds the best hyper plane that separates the data points of one class from the data points of another class.

APPLY IMPROVED SUPPORT VECTOR MACHINE ALGORITHM AND EVALUATE ACCURACY:

Support vector machine algorithm is improved by introducing specific kernel features and then the accuracy is calculated.

COMPARATIVE ANALYSIS OF THE ALGORITHMS IN TERMS OF ACCURACY, PRECESION AND RECALL:

Comparative analysis of the entire algorithms is done and the result of performance is calculated in terms of accuracy, precision and recall.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The objective of this proposed work is to have greater accuracy, as high precision and recall metrics.

PRECISION: precision is how many selected items are relevant. It is a ratio of true positive to the sum of true positive and false positive.

RECALL: recall is how many relevant items are selected. It is a ratio of true positive to the sum of true positive and false negative.

For the implementation of our proposed algorithm I have used Matlab version 2015, with i7 processor with ram of 8gb having processor speed 2.7ghz, for fast optimization of our algorithm we initialized Matlab pool using the 'local' profile, nevertheless as we can see that the output performance of various classifier in figure the performance of improved SVM (support vector machine) Outperformed the rest of the classifier as it give the 100% correct classification rate for no, and 85.4% correct classification rate for yes, and over all classification rate is 94.8% which is the highest accuracy achieved present in the literature.

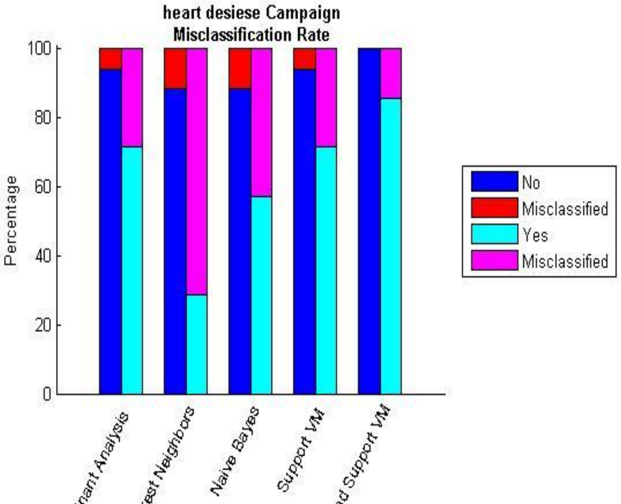


***Table 1** Prediction accuracy between various various data minnig techniques*

Data mining technique	Precision	Recall	Accuracy
K-mean based MAFIA	0.78	0.64	0.74
K-mean based MAFIA with ID3	0.8	0.84	0.84
K-mean based MAFIA with ID3 and C4.5	0.82	0.89	0.89
Proposed	0.92	0.94	0.948

VI. CONCLUSION AND FUTURE WORK

Medical related information is highly voluminous in nature in the healthcare industry. It can be derived or retrieved from various sources which are not entirely applicable in this feature. In this work, heart disease prediction system was developed using classification algorithms through Matlab data mining tool to predict effective and better accurate results regarding whether the patient is suffering from heart disease or not. In future work, we have planned to propose more effective heart disease prediction system to predict heart diseases with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms.

REFERENCES

[1] Abhishektaneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[2] Aqueel Ahmed, Shaikh Abdul Hannan,”Data Mining Techniques to Find Out Heart Diseases”, International Journal of Innovative Technology and Exploring Engineering(IJITEE)ISSN: 2278-3075, Volume-1,Issue-4,September 2012.

[3] Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, “A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology”,2012

[4] Chitra R and Seenivasagam V, “REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES”, ISSN: 2229-6956(ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, JULY 2013, VOLUME: 03, ISSUE: 04, 2013

[5] Ms. Ishtake S.H, Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, 2013

[6] Nidhi Bhatla and Kiran Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol. 1 Issue 8, October – 2012

[7] M.A.Nishara Banu and B.Gomathy,” Disease Forecasting System Using Data Mining Methods”,2014

[8] Rashedur M. Rahman, FarhanaAfroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[9] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012