Decision Support System for Heart Disease Prediction using Data Mining Classification Techniques

Ankur Makwana
Department of Computer Science and Engineerin
Nirma University

Jaymin Patel
Department of Computer Science and Engineering
Nirma University

ABSTRACT

Data Mining techniques have been widely used to mine knowledgeable information from medical database. Most nations face high and expanding rates of coronary illness or Cardiovascular Disease. Despite the fact that, current pharmaceutical is creating colossal measure of information consistently, little has been carried out to utilize this accessible information to illuminate the difficulties that face a beneficial understanding of electrocardiography examination results. Computer situated in development alongside creditable Data Mining systems are utilized for proper results. Disease finding is one of the applications where Data Mining devices are demonstrating successful results. These are the main reason for death everywhere throughout the world in the past ten years. Several scientists are utilizing factual and Data Mining apparatuses to over assistance social insurance experts in the analysis of these disease. Using Hybrid Data Mining strategy in the analysis of coronary illness has been completely explored indicating satisfactory levels of accuracy.

Keywords:

Data Mining; Decision Support System; Health care;Health records; Classification

1. INTRODUCTION

Data Mining is a crucial step in discovery of knowledge from large data sets. In recent years, Data Mining has found its significant hold in every field including health care. Data Mining techniques can be used for data selection, finding patterns and predict the diseases using large data. Mining process is more than the data analysis which includes classification, clustering, and association rule discovery[1]. It also spans other disciplines like Data Warehousing, Statistics, Machine learning and Artificial Intelligence. Data Mining is predicted to be one of the most revolutionary developments of the century, according to the online technology magazine ZDNET News. In fact, The Technology Review (S. Apte, 2012) chose Data Mining as one of 10 emerging technologies that will change the world. Effectively analyzing information from customers, partners, and suppliers has become important to more companies[1].
Furthermore, many companies have implemented a data warehouse strategy and are now starting to look at what they can do with all that data (R. Awang, 2008). Data Mining can be a useful tool in the health sector and health care. Organizations that perform Data Mining are better positioned to meet their long-term needs (A. Khemphila and V. Boonjing, 2007) argue that data can be a great asset to health care organizations, but they have to be first transformed into information[4]. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop Data Mining applications. As of late new research avenues, for example, naive Bayes, which incorporates Data Mining techniques, has turned into a famous examination tool for medical analysts who look for to recognize and utilize examples and connections among huge number of variables, and have the capacity to anticipate the result of a disease utilizing the historical cases put away inside data sets.

1.1 Definition

Nowadays, Health care industry contains huge amount of heath care data and these health care data contains hidden information. This hidden information is useful for making effective decisions using different Data Mining techniques. We can develop an efficient decision making for patients who will be suffering from disease and using this we can identify patient for improve their health and early prediction.

1.2 Objective of Study

As shown in the Fig-1 identify the patients who will be suffering from disease using Data sets. To use these data sets find out the different techniques for better efficiency and accuracy of the prediction. Hybrid Data Mining technique will be used to reduce cost and best outcomes, to improve health.

—To identify key patterns or features from the data set.
—To Identify and select attributes that are more relevant in relation to Heart Diseases.

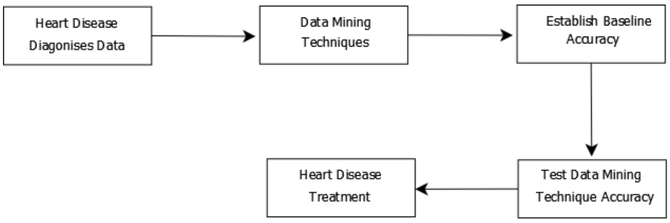
Heart Disease Diagonises Data
Data Mining Techniques
Establish Baseline Accuracy
Test Data Mining Technique Accuracy
Heart Disease Treatment

Fig. 1. complete scenario of identify the Heart disease patients.

2. LITERATURE SURVEY

Work done in patients identification using historical data sets using different algorithm and techniques are discussed below:

Coronary illness is the main reason for death on the planet in the course of recent years (World Health Organization 2007). The European Public Health Alliance reported that heart assaults, strokes and other circulatory sicknesses represent 41% of all passing (European Public Health Alliance 2010)[14].
A few different symptoms are connected with coronary illness, which makes it hard to diagnose it faster and better. Dealing with coronary illness patients databases can be contrasted with genuine application. Specialist's information to appoint the weight to each one property. More weight is allotted to the quality having high effect on sickness forecast. In this manner it seems sensible to take a stab at using the learning and knowledge of a few pros gathered in databases towards aiding the conclusion process. It additionally gives social insurance experts an additional wellspring of information for deciding[7].
The medicinal services industry gathers a lot of social insurance information and that need to be mined to find concealed data for compelling choice making. Propelled by the overall expanding mortality of coronary illness patients every year and the accessibility of colossal measure of patient's information from which to concentrate valuable learning, analysts have been utilizing Data Mining strategies to help medicinal services experts in the determination of coronary illness[1].
Forecast of coronary illness utilizing Data Mining systems has been a continuous exertion for as far back as two decades. The vast majority of the papers have actualized a few information digging systems for determination of coronary illness, for example, Decision Tree, Naive Bayes, Neural System, portion thickness, consequently characterized gatherings, sacking calculation and help vector machine demonstrating diverse levels of correctness's on numerous databases of patients from around the globe[7].
One of the bases on which the papers differ are the determination of parameters on which the methods have been used. Various makers have shown differing parameters also, databases for testing the precision.Specifically, professionals have been exploring the utilization of the Decision Tree system in the finding of coronary infection with foremost achievement.Used the R-programming to investigate applying Naive Bayes and Decision Trees for the area of coronary illness.Utilized the pressing count as a piece of the R-programming contraption and differentiated it and Decision Tree in the investigation of coronary illness.In the choice making arrangement of coronary ailment is successfully diagnosed by Random Forest number. In concentrated around the likelihood of choice sponsorship, the coronary disorder is predicted. Along these lines the originator considered that decision tree performs well and every now and again the precision is practically identical in Bayesian request[8].
C. S. Dangare and S. S. Apte, 2012 performed a work, "Improved study of heart disease prediction system using data mining classification techniques"[1].In this paper discuss the grouping tree methods in Data Mining. The grouping tree calculations utilized Tree calculation. The goal of this examination was to analyze the results of the execution of diverse arrangement methods for a coronary illness data set[3].
Data Mining is the investigation of substantial data sets to concentrate shrouded and awhile ago obscure examples, connections and learning that are hard to distinguish with customary factual techniques (A. Khemphila and V. Boonjing ,2010). Hence Data Mining alludes to mining or concentrating learning from a lot of

Literature survey Table:

Author	Year	Techniques	Suggestion
Yuehjen E. Shao,Chia-Ding Hou,Chih-Chou Chiub [11]	2014	Hybrid intelligent modeling schemes	There should be techniques for finding missing value like admission,emergency to better utilized for predict risk
Rashedur M. Rahman,Fazle Rabbi Md.Hasan [10]	2014	Neural network for classification	Used Machine Learning techniques for Diseases
M.Akhil jabbar,B.L Deekshatulu,Priti Chandra [3]	2013	propose a new algorithm which combines KNN with Genetic Algorithm for effective classification.	Give various weights to the presence of condition for other disease
Jesmin Nahar,Tasadduq Imam,Kevin S.Tickle [8]	2013	Support Vector Machine Algorithm	Apply Support Vector Machine for automated feature selection and a medical knowledge based motivated feature selection process.
Jesmin Nahar, Kevin S.Tickle, Yi-Ping Phoebe Chen [7]	2013	Rule Mining to detect factors which contribute to heart disease	It should be deal with missing value
Dangare, Chaitrali S and Apte [1]	2012	Use Classification Techniques for Improved Study of Heart Disease Prediction System	It can use more number of attributes so get more accurate results.
S. Muthukaruppan, M.J. Er [5]	2010	fuzzy expert system	For use only longitudinal data set it can not increase predict power of data.
Anchana Khemphila, Veera Boonjing [4]	2010	Logistic regression classifiers , Artificial neural networks , Classification trees , Decision trees	Comparing performances for classifying heart disease patients
Hongmei Yan, Jun Zheng, Yingtao Jiang [14]	2008	Genetic Algorithm	Not depends on only one at- tribute for accurate prediction.
Sellappan Palaniappan, Rafiah Awang [9]	2008	Decision Trees, Nave Bayes and Neural Network algorithm	system extracts hidden knowledge from a historical heart disease database

information[4]. Data Mining applications will be utilized for better well being strategy making and counteractive action of doctor's facility lapses, early discovery, anticipation of infections and preventable healing center passing (P. Chandra 2013)[2]. Coronary illness expectation framework can support therapeutic experts in foreseeing coronary illness focused around the clinical information of patients. Thus by actualizing a coronary illness forecast framework utilizing Data Mining procedures and doing an Data Mining on different coronary illness properties, it can ready to anticipate all the more probabilistic-ally that the patients will be diagnosed with coronary illness.

2.1 Research design

To give an examination among the well mainstream characterization calculations, four execution measurements were utilized as a part of our analysis. These are exactness, genuine positive rate (TP), F-measure, and time. Here, precision was the general forecast exactness, genuine positive rate (TP) was the precise grouping rate for the positive classes, and F-measure demonstrates the viability of a calculation at the point when the precise forecast rates for both of the classes are considered. Additionally, preparing time was considered to analyze the computational unpredictability for learning[1]. On account of therapeutic information finding, numerous scientists have utilized a 10-fold cross acceptance on the aggregate information and reported the result for sickness identification, while different scientists have not utilized this system for coronary illness expectation[1]. We contend that selecting the best preparing parameters on an approval set and reporting forecast on a test set is more legitimate than just performing a 10-fold cross approval on a preparation set. Be that as it may, to relate with regular society, we have utilized both the train-test part strategy and 10-fold cross acceptance when contrasting the calculations[6].

3. STUDY AND COMPARISON OF AVAILABLE TECHNIQUES

3.1 Genetic Algorithm :

Evolutionary processing began by lifting thoughts from organic hypothesis into software engineering. Genetic Algorithm are most mainstream method in evolutionary registering. Evolutionary calculations are utilized as a part of issues for enhancement. To take care of issues, evolutionary algorithm oblige an information structure to speak to and assess arrangement from old arrangements.
Genetic Algorithm (GA) was created by John Holland in 1975. Genetic calculations are helpful for hunt and streamlining problems. GA utilizes hereditary qualities as its model as critical thinking. Every arrangement in Genetic Algorithm is spoken to through chromosomes. Chromosomes are comprised of qualities, which are individual components that speak to the issue[3]. The accumulation of all chromosomes is called populace. For the most part there are three prominent administrators are use in GA[3].

—Selection : This operator is used in selecting individuals for reproduction.

—Crossover : This is the methodology of taking two guardian chromosomes and creating a child from them.

—Mutation : This administrator is utilized to modify the new arrangements in the quest for better arrangement. Change keeps the GA to be caught in a local minimum.

—Fitness function : wellness work in GA is the estimation of a target capacity for its phenotype. The chromosome must be initially decoded, for calculating the fitness function[3].

3.2 Naive Bayes Classifier :

Naive Bayesian classifiers have turned out to be effective apparatuses for tackling order issues in a mixture of spaces. A Naive Bayesian classifier, fundamentally, is a model of a joint likelihood appropriation over a set of stochastic variables[10]. It is made out of a solitary class variable, demonstrating the conceivable conclusions or classes for the issue under study, and a set of peculiarity variables, displaying the peculiarities that accommodate recognizing between the different classes; the gimmick variables are thought to be commonly free given the class variable. Cases of the order issue under study are displayed to the classifier as a mix of values for the gimmick variables; the classifier then returns a back likelihood dispersion over the class variable, that is, it yields a probabilistic rundown for each one class.Credulous Bayesian classifier have been effectively connected in the medicinal area where they are utilized for comprehending symptomatic issues. Naive Bayesian classifiers are regularly gained from information. Adapting such a classifier sums to securing the earlier probabilities of the diverse classes and assessing the contingent probabilities of the different gimmicks given each of the classes[10].

3.3 Decision trees algorithm :

Decision tree was considered here among different varieties of Data Mining procedures because of these underneath reasons.Decision tree channels are not difficult to actualize and straightforward. It is a system usually utilized as a part of Data Mining. Decision tree is one of the Data Mining systems demonstrating extensive achievement when contrasted with other Data Mining methods[4]. It is a choice emotionally supportive network that uses a tree-like chart choices. Decision trees are the most influential methodologies in learning disclosure and Data Mining. Decision trees are exceedingly powerful instruments in numerous zones, for example, information and content mining, data extraction, machine learning,and example disestablishment. It can deal with information like Nominal, Numeric and Text[4]. It has the capacity process mistaken data sets or missing qualities.
A Decision Tree is utilized to take in an arrangement capacity which closes the estimation of a ward property (variable) given the estimations of the autonomous (info) characteristics. This confirms an issue known as administered grouping in light of the fact that the ward characteristic and the checking of classes (qualities) are given[7]. Tree multifaceted nature has its impact on its exactness. Normally the tree many- sided quality can be measured by a measurements that contains: the aggregate number of hubs, aggregate number of leaves, profundity of tree and number of properties utilized as a part of tree development[6].

3.4 Logistic Model Tree Algorithm :

Logistic Model Tree is the classifier for building 'logistic model trees', which comprise of a choice tree structure with logistic relapse capacity at the clears out. The calculation can manage twofold and multiple class target variables, numeric and ostensible qualities and missing qualities[4]. A mix of learners that depend on basic relapse models if little and/or loud information is accessible and include a more unpredictable tree structure if there is sufficient information to warrant such a structure. LMT uses cost-unpredictability pruning. This calculation is altogether slower than alternate algorithm[13].

3.5 Random forest algorithm :

Ramdom Forest is an outfit classifier that incorporates different choice trees. The yield of the classes is tended to by individual trees. It is gotten from unusual choice a timberland that was proposed by Tin Kam Ho of Bell Labs in 1995[10]. This system hardens with self-self-assured determination of qualities to add to a choice trees with controlled blended packs. The tree is made utilizing estimation as examined
The Random Forest procedure has some alluring qualities, for example,

- —It is not hard to use, fundamental and easy parallelism.
- —It doesn't oblige models or parameters to pick beside the amount of markers to pick at self-assertive at each center[3].
- —It runs effectively on extensive databases; it is moderately strong to anomalies and commotion.
- —It can deal with a huge number of information variables without variable erasure; it gives evaluations of what variables are paramount in the arrangement.
- —It has a successful system for assessing missing information and keeps up precision when a vast extent of the information are missing, it has routines for adjusting lapse in class populace unequal information[10].

4. PROPOSED FRAMEWORK

4.1 Proposed method

Our proposed approach combines Naive Bayes and Genetic Algorithm to improve the classification accuracy of heart disease data set. We utilized hereditary inquiry as a decency measure to prune excess and superfluous characteristics, and to rank the characteristics which help more towards order.Slightest positioned characteristics are uprooted, and order calculation is based focused around assessed characteristics[12]. This classifier is prepared to order coronary illness information set as either sound or debilitated. Our proposed algorithm comprises of two sections.

- —As shown in the Fig-2 To start with part manages attributes using genetic search.
- —Second part deals with building classifier and measuring accuracy of the classifier Proposed algorithm.

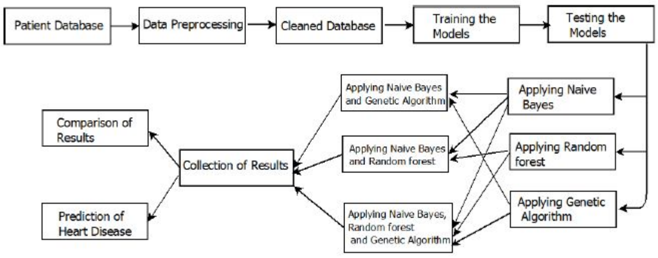
Patient Database
Data Preprocessing
Cleaned Database
Training the Models
Testing the Models
Applying Naive Bayes and Genetic Algorithm
Applying Naive Bayes
Comparison of Results
Applying Naive Bayes and Random forest
Applying Random forest
Collection of Results
Applying Genetic Algorithm
Prediction of Heart Disease
Applying Naive Bayes, Random forest and Genetic Algorithm

Fig. 2. Proposed Framework

4.1.1 Patient Databaset. Patient database is data sets collected from Cleveland Heart Disease Data set (CHDD) available on the UCI Repository.

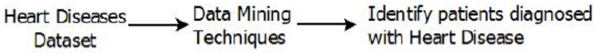

Fig. 3. Flow of the Work

4.1.2 Data Preprocessing. This phase includes extraction of data from the Cleveland Heart Disease Data set (CHDD) in a uniform format.

4.1.3 Training the Models. Each of the three models has been trained using different methods.

4.1.4 Testing the Models. This type of model used to analyze data and discover patters in classification and regression analysis.

4.1.5 Comparison of Results. The results obtained after applying the rules will be analyzer on the basis of sensitivity, specificity, and accuracy.

5. CONCLUSION

In this study, the aim is to design a predictive model for heart disease detection using Machine Learning and Data Mining techniques. Data Mining techniques can be used efficiently to model and predict heart disease using Naive Bayes and Genetic Algorithm provides accurate results. The fuzzy relapse model was found to have the capacity of evaluating the connections between the input of predicted patients and predicted outcomes of patients results. This methodology gives a decent answer for manage instability in health framework variables and instability in the admission of a patient.

6. FUTURE WORK

The technique can be improved further by experiments with more data set and using hybrid algorithms to improve the classification accuracy and to build a model that can predict specific heart disease. In order to show a good prediction system we can not adjust the prediction system.We are currently investigation the methods to make better use of this information. To improve the accuracy we have to build a model that can predict heart Disease.

7. REFERENCES

[1] Chaitrali S Dangare and Sulabha S Apte. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10):44–48, 2012.

[2] Mu-Jung Huang, Mu-Yen Chen, and Show-Chin Lee. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, 32(3):856–867, 2007.

[3] M Akhil Jabbar, BL Deekshatulu, and Priti Chandra. Heart disease classification using nearest neighbor classifier with feature subset selection. *Anale. Seria Informatica*, 11, 2013.

[4] Anchana Khemphila and Veera Boonjing. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. In *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, pages 193–198, 2010.

[5] S Muthukaruppan and Meng Joo Er. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. *Expert Systems with Applications*, 39(14):11657–11665, 2012.

[6] T Mythili, Dev Mukherji, Nikita Padalia, and Abhiram Naidu. A heart disease prediction model using svm-decision trees-logistic regression (sdl). *International Journal of Computer Applications*, 68(16):11–15, 2013.

[7] Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4):1086–1093, 2013.

[8] Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1):96–104, 2013.

[9] Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE, 2008.

[10] Rashedur M Rahman and Fazle Rabbi Md Hasan. Using and comparing different decision tree classification techniques for mining icddr, b hospital surveillance data. *Expert Systems with Applications*, 38(9):11421–11436, 2011.

[11] Yuehjen E Shao, Chia-Ding Hou, and Chih-Chou Chiu. Hybrid intelligent modeling schemes for heart disease classification. *Applied Soft Computing*, 14:47–52, 2014.

[12] Mai Shouman, Tim Turner, and Rob Stocker. Using data mining techniques in heart disease diagnosis and treatment. In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*, pages 173–177. IEEE, 2012.

[13] Yanwei Xing, Jie Wang, Zhihong Zhao, and Yonghong Gao. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *Convergence Information Technology, 2007. International Conference on*, pages 868–872. IEEE, 2007.

[14] Hongmei Yan, Jun Zheng, Yingtao Jiang, Chenglin Peng, and Shouzhong Xiao. Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. *Applied soft computing*, 8(2):1105–1111, 2008.