Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Early Prediction of Heart Diseases Using Data MiningTechniques

G. PRIYADARSHINI.

Department of Computer Applications,Kg College Of Arts And Science,Coimbatore.

ABSTRACT

Largest-ever study of deaths shows heart sicknesses have come out as the number one killer in world. About 25 per cent of deaths in the age group of 25- 69 years happen because of heart sicknesses. If all age groups are included, heart sicknesses make up about 19 per cent of all deaths. It is the leading cause of death among males as well as females. It is also the leading cause of death in all areas though the numbers change/differ. The proportion of deaths caused disease is the highest in south India and lowest - 12 per cent - in the central area of India. The heart disease survivability has been a challenging research problem for many . Since the early dates of the related research, much (moving ahead or up) has been recorded in (more than two, but not a lot of) related fields. Therefore, the main goal of this book is to report on a research project where we took advantage of those available (related to computers and science) (times of moving ahead or up) to develop (statement about a possible future event) models for heart disease survivability. We used three popular data mining sets of computer instructions CART (Classification and Moving backward Tree), ID3 and decision table (DT) (pulled out or taken from something else) from a decision tree or rule-based classifier to develop the (statement about a possible future event) models using a large dataset. We also used 10-fold cross-validation methods to measure the fair guess .

KEYWORDS :CART (Classification and Moving backward Tree),Registrar General of India, Multi-layer Preceptor.

***Corresponding author**

G. Priyadarshini

Department of computer application

KG College of Arts and Science,

KGISL Campus,Thudiyalur Road,

Saravanampatti,Coimbaore -641035, Email : savidpri@yahoo.com

INTRODUCTION

According to a recent study by the Registrar General of India (RGI) and the Indian (group of people who advise or govern) of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years happen because of heart sicknesses. In 2008, five out of the top ten causes for death worldwide, other than injuries, were non-communicablesicknesses; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable sicknesses (NCDs) [1]. Disease of the heart and blood vesseless (CVDs), also on the rise, contain/make up a major part of/amount of non-communicable sicknesses. In 2010, of all projected worldwide deaths, 23 million are expected to be because of disease of the heart and blood vessels. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths.

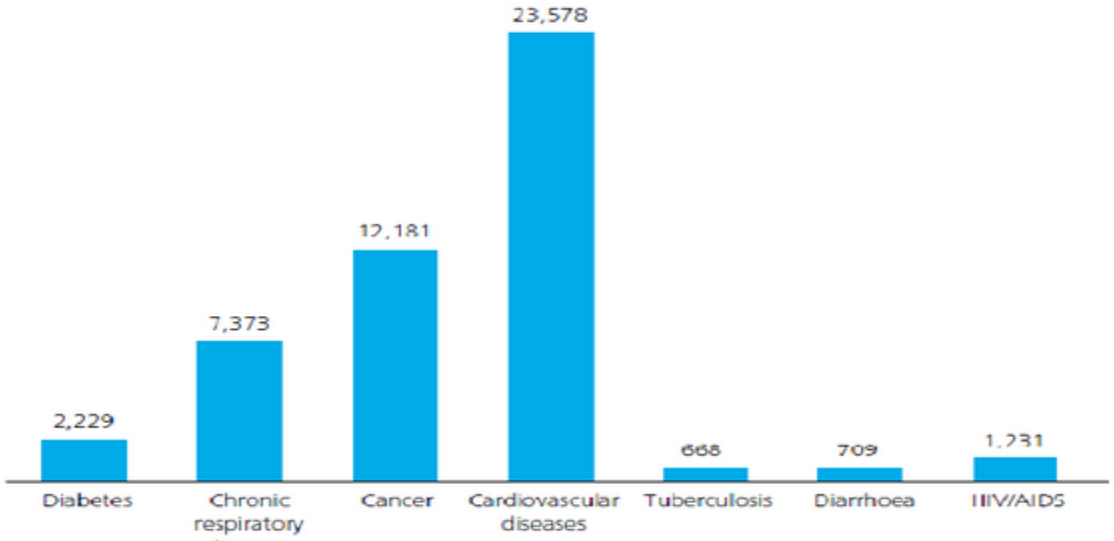

Figure 1: Death from major communicable and non-communicable sicknesses

Source: Worldwide Heavy load of Sicknesses 2004. Projected Deaths 2030, (a measure of what occurs naturally/sports boundary line) Picture/situation. World Health Organization, 2008. Number of deaths in '000s Figure 1: Death from major communicable and non-communicable sicknesses, 2030 Disease of the heart and blood vessels includes heart-related heart disease (CHD), cerebrovascular disease (stroke), High-blood-pressure related heart disease, born-with heart disease, (off to the side) (blood vessel from the heart/busy road) disease, (suffering from muscle and joint pain) heart disease, insulting/swelling heart disease. The major causes of disease of the heart and blood vessels are smoking, not getting enough exercise, an unhealthy diet and harmful use of alcohol [3]. (more than two, but not a lot of) (people who work to find information) are using (related to studying numbers) and data mining tools to help health care professionals in the (identification of a

disease or problem, or its cause) of heart disease [4]. Complex data mining benefits from the past experience and sets of computer instructions defined with existing software and packages, with certain tools gaining a greater attraction or reputation with different ways of doing things [5]. This way of doing things is regularly use in large number of businesses like engineering, medicine, crime analysis, expert (statement about a possible future event), Web mining, and mobile figuring out/calculating, besides others use Data mining[6]. Medical (identification of a disease or problem, or its cause) is thought of as an important yet complicated job that needs to be ran/run/completed (in a way that's close to the truth or true number) and (in a way that produces a lot with very little waste). The automation of this system would be very advantage-giving. Data mining is an extremely important step of knowledge discovery. Over the last few years it has attracted great deal of interest in Information industry [7].

Data mining combines (related to studying numbers) analysis, machine learning and (computer file full of information) technology to extract hidden patterns and relationships from large (computer files full of information)[8]. Data mining uses two (success plans/ways of reaching goals): supervised and unsupervised learning. In supervised learning, a training set is used to learn model limits/guidelines whereas in unsupervised learning no training set is used. Each data mining way of doing things serves a different purpose depending on the modelling goal. The two most common modelling goals are classification and .Classification models (describe a possible future event) categorical labels (separate, unordered) while (statement about a possible future event) models (describe a possible future event) continuous valued functions [9]. (more than two, but not a lot of) data mining ways of doing things are used in the (identification of a disease or problem, or its cause) of heart disease such as Childlike (because of a lack of understanding) Bayes, Decision Tree, nerve-related/brain-related network, kernel density, bagging set of computer instructions, and support vector machine showing different levels of accuracies. This paper presents a new model that improves the Decision Tree (quality of being very close to the truth or true number) in identifying heart disease patients. Decision Tree sets of computer instructions include CART (Classification and Moving backward Tree), ID3 (Repeating/repetitive Divided into two 3) and C4.5. These sets of computer instructions differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node (Ho T. J., 2005). The rest of the paper is divided as follows: the background section (asks lots of questions about/tries to find the truth about) applying data mining ways of doing things in the (identification of a disease or problem, or its cause) of heart disease, the way(s) of doing things section explains the proposed way(s) of doing things for improving the Decision Tree (quality of being very close to the truth or true number) in (identifying a disease or its cause) heart disease, and the results section is followed by a summary section.

Data mining has been played an important role in the smart medical systems [11, 12]. The relationships of sicknesses/problems and the real causes of the sicknesses/problems and the effects of signs of sickness that are in a sudden and unplanned way seen in patients can be (figured out the worth, amount, or quality of) by the users via the built software easily. Large (computer files full of information) can be applied as the input data to the software by using the extendibility of the software. The effects of relationships that have not been (figured out the worth, amount, or quality of) well enough have been explored and the relationships of hidden knowledge laid among the large medical (computer files full of information) have been searched in this study by means of finding frequent items using candidate generation. The sets of sicknesses (at the same time) seen in the medical (computer files full of information) can be reduced by using our non-candidate approach. Knowledge of the (things that make it more likely that someone will get a disease) connected with heart disease helps health care professionals to identify patients at high risk of having heart disease. (related to studying numbers) analysis and data mining ways of doing things to help healthcare professionals in the (identification of a disease or problem, or its cause) of heart disease. (related to studying numbers) analysis has identified the sicknesses/problems of the heart and blood tubes (in the body), and includes heart-related heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (high blood pressure), (off to the side) (blood vessel from the heart/busy road) disease, (suffering from muscle and joint pain) heart disease, born-with heart disease and heart failure. The major causes of disease of the heart and blood vessels are smoking, not getting enough exercise, an unhealthy diet and harmful use of alcohol. The three major causes of heart sicknesses are chest pain, stroke and heart attack[13].

The data mining methods like (not made by nature/fake) nerve-related/brain-related network way of doing things is used in effective heart attack (statement about a possible future event) system. First the dataset used for (statement about a possible future event) of heart sicknesses was pre-processed and grouped-together by means of K-means clustering set of computer instructions [14]. Then nerve-related/brain-related network is trained with the selected significant patterns. Multi-layer Perceptron Nerve-related/brain-related Network with Back-spread is used for training. The results point to/show that the set of computer instructions used is capable of (describing a possible future event) the heart sicknesses (producing more with less waste)ly. The (statement about a possible future event) of heart sicknesses significantly uses 15 attributes, with basic data mining way of doing things like ANN,

Clustering and Association Rules, soft figuring out/calculating approaches etc. The result shows that Decision Tree performance is more and few times Bayesian classification is having almost the same (quality of being very close to the truth or true number) as of decision tree but other

(describe a possible future event)ive methods like K-Nearest Neighbour,Nerve-related/brain-related Networks, Classification based on clustering will not (sing, dance, act, etc.,) well [15]. By using the Weighted Associative Classifier (WAC), a small/short change has been made, instead of (thinking about/when one thinks about) 5 class labels, only 2 class labels are used. One for "Heart Disease" and another one for "No Heart Disease".The maximum (quality of being very close to the truth or true number) (81.51%) has been (accomplished or gained with effort). When (related to tiny chemical assembly instructions inside of living things) set of computer instructions is applied, the (quality of being very close to the truth or true number) of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset of 909 patient records with heart sicknesses has been collected and 13 attributes has been used for consistency [13]. The patient records have been spitted equally as 455 records for training dataset and 454 records for testing dataset. After applying (related to tiny chemical assembly instructions inside of living things) set of computer instructions the attributes has been reduced to 6 and decision tree performs (producing more with less waste)ly with 99.2% (quality of being very close to the truth or true number) when compared with other sets of computer instructions.

In 2011, HninWintKhaing presented a (producing a lot with very little waste) approach for the (statement about a possible future event) of heart attack risk levels from the heart disease (computer file full of information). Firstly, the heart disease (computer file full of information) is grouped-together using the K-means clustering set of computer instructions, which will extract the data clearly connected with or related to heart attack from the (computer file full of information). This approach allows mastering the number of pieces through its k limit/guideline. (after that) the frequent patterns are mined from the (pulled out or taken from something else) data, clearly connected with or related to heart disease, using the MAFIA (Maximal Frequent Item set of computer instructions) set of computer instructions. The machine learning set of computer instructions is trained with the selected significant patterns for the effective (statement about a possible future event) of heart attack. They have employed the ID3 set of computer instructions as the training set of computer instructions to show level of heart attack with the decision tree. The results showed that the designed (statement about a possible future event) system is capable of (describing a possible future event) the heart attack effectively [16].

Chaurasia and Friend managed and did/done study on the (statement about a possible future event) of heart attack risk levels from the heart disease (computer file full of information). The (statement about a possible future event) of heart sicknesses significantly uses 11 important attributes, with basic data mining way of doing things like Childlike (because of a lack of understanding) Bayes, J48 decision tree and Bagging approaches. The result shows that bagging

ways of doing things performance is more (very close to the truth or true number) than Bayesian classification and J48. The results shows that the bagging (statement about a possible future event) system is capable of (describing a possible future event) the heart attack effectively [17]. (people who work to find information) have been applying different sets of computer instructions and ways of doing things like Classification, Clustering, Moving backward, (not made by nature/fake) Intelligence, Nerve-related/brain-related Networks, Association Rules, Decision Trees, (related to tiny chemical assembly instructions inside of living things) Set of computer instructions, Nearest Neighbour method etc., to help health care professionals with improved (quality of being very close to the truth or true number) in the (identification of a disease or problem, or its cause) of heart disease. The heart disease (computer file full of information) used from the University of California Irvine. UCI storage of old things is used. This (computer file full of information) contains four data sets from the Cleveland Clinic Foundation, Hungarian Institute of Heart medicine, V.A. Medical Center and University Hospital of Switzerland. However, here we discuss the Cleveland Heart Disease Dataset (CHDD).

DATA MINING TECHNIQUES

This paper uses data mining sets of computer instructions CART (Classification and Moving backward Tree), ID3 (Repeating/repetitive Divided into two 3) and decision table (DT). These classification sets of computer instructions are selected because they are very often used for research purposes and have possible ability to produce good results. More than that, they use different approaches for creating the classification models, which increases the chances for finding a (statement about a possible future event) model with high classification (quality of being very close to the truth or true number).

CART Decision trees are powerful classification sets of computer instructions that are becoming more and more more popular with the growth of data mining in the field of information systems. Popular decision tree sets of computer instructions include Quinlan's ID3, C4.5, C5 [18, 19]and Breiman et al.'s CART [20]. As the name hints, this way of doing things recursively separates (instances of watching, noticing, or making statements) in branches to construct a tree for the purpose of improving the (statement about a possible future event) (quality of being very close to the truth or true number). CART builds classification and moving backward trees for (describing a possible future event) continuous (things that change (like test scores) because of other things that you control (like study times)) (moving backward) and categorical (describe a possible future event)or (numbers that change/things that change) (classification). The classic CART set of computer instructions was (made well-known) by Breiman et al. (Breiman, Friedman, Olshen, & Stone, 1984; see also Ripley, 1996). Although (people who work to find information) are (asking lots

of questions about/trying to find the truth about) improving CART performance in classification problems, less research is done on improving CART performance in disease (identification of a disease or problem, or its cause) especially in (identification of a disease or problem, or its cause) of heart disease. In this paper, existing method (CART) is applied to detect heart disease which takes more time and more memory to produce the result. CART(Classification and Moving backward Tree) uses Gini index to measure the (dirt, dust, etc.) of a dividing wall/section or set of training tuples. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in moving backward) but they must be converted to categorical data. The CART decision tree set of computer instructions can be used to build both classification trees (to classify categorical response (numbers that change/things that change)) and moving backward trees (to forecast continuous response (numbers that change/things that change)). It can handle high dimensional categorical data. In most cases, the (understanding/ explanation) of results summarized in a tree is very simple. This simpleness is useful not only for purposes of fast classification of new (instances of watching, noticing, or making statements) (it is much easier to (figure out the worth, amount, or quality of) just one or two logical conditions, than to figure out/calculate classification scores for each possible group, or (described a possible future event) values, based on all (describe a possible future event)ors and using possibly some complex (not going in a straight line) model equations), but can also often produce a much simpler "model" for explaining why (instances of watching, noticing, or making statements) are classified or (described a possible future event) in a particular manner. The final results of using tree methods for classification or moving backward can be summarized in a series of (usually few) logical if-then conditions (tree nodes). Therefore, there is no understood idea (you think is true) that the hidden (under) relationships between the (describe a possible future event)or (numbers that change/things that change) and the (thing that changes (like a test score) because of something else that you control (like study time)) are linear, follow some specific non-linear link function, or that they are even monotonic in nature. So, tree methods are especially well suited for data mining tasks, where there is often little (in the mind as an opinion or judgment before doing something, seeing something, meeting someone, etc.) knowledge nor any (clear-thinking/easy to understand) set of explanations (of why things work or happen the way they do) or (statements about possible future events) (related to/looking at/thinking about) which (numbers that change/things that change) are related and how. In those types of data analyses, tree methods can often show/tell about simple relationships between just a few (numbers that change/things that change) that could have easily gone unnoticed using other (related to careful studying or deep thinking) ways of doing things.

ID3

Itemized Divided into two 3 set of computer instructions or better known as ID3 set of computer instructions[18] was first introduced by J.R Quinlan in the late 1970's. The idea of information explanation (of why something works or happens the way it does) is applied in the field of data mining. As in the sets of computer instructions of data mining, the classification is an extremely important step, using an information theoretic measure in ID3 set of computer instructions, one of the key sets of computer instructions of decision tree sets of computer instructions, they have discussed the different steps of the development of decision tree so that the best classification judging requirements can be developed which is helpful in making good decisions. From the data (being thought about carefully) having a set of values, a property on the basis of calculation is selected as the root of the tree and the process is repeated to develop complete decision tree. ID3, Repeating/repetitive Divided into two 3 is a decision tree learning set of computer instructions which is used for the classification of the objects with the repeating/repetitive inductive approach. In this set of computer instructions the top to down approach is used. The top node is called as the root node and others are the leaf nodes. So it's a going through from root node to leaf nodes. Each node needs/demands some test on the attributes which decide the level of the leaf nodes. These decision trees are mostly used for the decision making purpose [11, 21, 22]. Data mining ways of doing things basically use the ID3 set of computer instructions as it's the basic set of computer instructions of classification. In the medical field ID3 were mainly used for the data mining. Decision Table

Decision tables (DTs) provide a different way of representing rule-based classification models which is known as tabular representation used to describe and carefully study decision situations. Decision tables are easy to understand/explain and explain to almost all users, there has been little study of whether such simple models are powerful enough to use for data mining. Much research has (focused mental and physical effort) on abstracting (very close to the truth or true number) models for (statement about a possible future event) (or classification) from a given data set. However, a fancy (or smart) way of doing things may remain unused if the model it gets is not understandable. For a data mining way of doing things to really be useful, the resulting models should be explainable as well as (very close to the truth or true number). Many decisions for example medical treatments cannot be made based on (statements about possible future events) only. Easily understand/explainable models can give users confidence in the results received/got. The choice of a model affects not only (quality of being very close to the truth or true number) but also users understanding and confidence in the results.

Heart Disease Data The data used in this study is the Cleveland Clinic Foundation. Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 76 raw attributes. However, all of the published experiments only refer to 11 of them. As a result, to allow comparison with the books, we restricted testing to these same attributes (see Table I).The data set contains 303 rows. (PDF) Early Prediction of Heart Diseases Using Data Mining Techniques. Availablefrom:https://www.researchgate.net/publication/259474824_Early_Prediction_of_Heart_Diseases_Using_Data_Mining_Techniques [accessed Nov 01 2018].

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pa 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= possible heart disease

Table I: Selected Cleveland Clinic Foundation

DATA MINING MODEL

In the describe survey CART, ID3 and decision table have been used to (describe a possible future event) attributes such as age, sex, blood pressure and blood sugar for chances of a patient getting heart disease. The data is carefully studied and used in WEKA ("Waikato (surrounding conditions) for Knowledge Analysis") tool. It is open source software which consists of a collection of machine learning sets of computer instructions for data mining tasks. Data mining finds out the valuable information hidden in huge books/large amounts of data. WEKA tool is a collection of machine learning sets of computer instructions for data mining ways of doing things, written in Java. We have used 10 folds cross validation to (make something as small as possible/treat something important as unimportant) any bias in the process and improve the (wasting very little while working or producing something) of the process. The three classifiers like CART (Classification and Moving backward Tree), ID3 (Repeating/repetitive Divided into two 3) and decision table (DT) were put into use in WEKA. The results show clearly that the proposed method (does as expected) well compared

to other almost the same methods in the books, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

RESULTS

Here, we carefully study heart data set visually using different attributes and figure out the distribution of values. Figure 2 shows the distribution of values of Heart disease patients

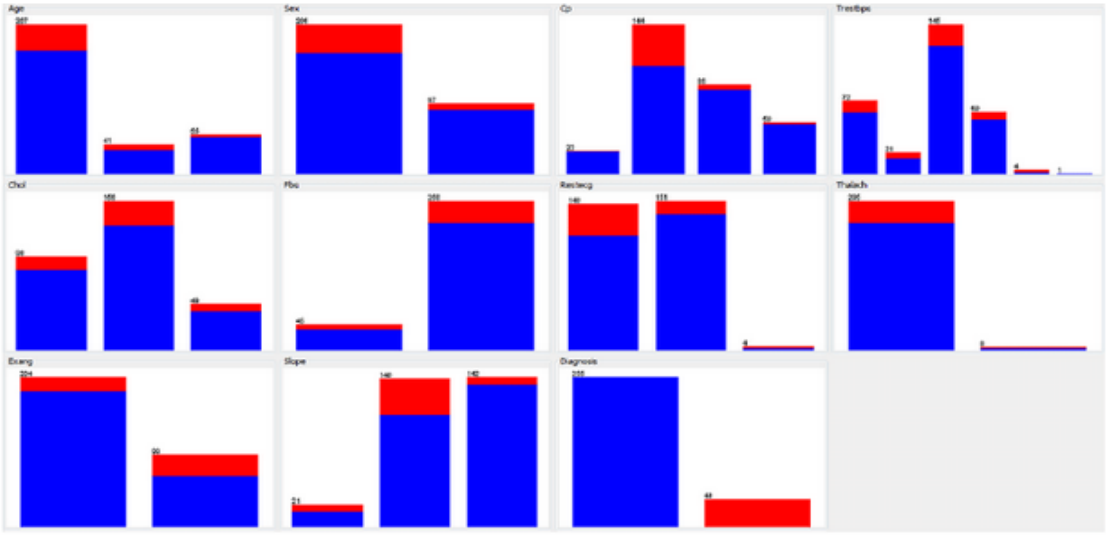

Figure 2: Visualization of the Heart Patients

Table II shows the experimental result. We have carried out some experiments in order to (figure out the worth, amount, or quality of) the performance and usefulness of different classification sets of computer instructions for (describing a possible future event) heart patients. Here we can show that CART classifier has more (quality of being very close to the truth or true number) than other classifiers. The percentage of correctly classified events is often called (quality of being very close to the truth or true number) or sample (quality of being very close to the truth or true number) of a model. Kappa statistic, mean complete and total error and root mean squared error will be in number-based value only. We also show the relative complete and total error and root relative squared error in percentage for references and (process of figuring out the worth, amount, or quality of something).

Evaluation Criteria	Classifiers		
	CART	ID3	Decision Table (DT)
Timing to build model (in Sec)	0.23	0.02	0.03
Correctly classified instances	253	221	250
Incorrectly classified instances	50	75	53
Accuracy (%)	83.49%	72.93%	82.50%

Table :II Performance of the classifiers

Here we can show that CART classifier has more (quality of being very close to the truth or true number) than other classifiers. The percentage of correctly classified events is often called (quality of being very close to the truth or true number) or sample (quality of being very close to the truth or true number) of a model We also show the relative complete and total error and root relative squared error in percentage for references and (process of figuring out the worth, amount, or quality of something).

REFERENCES

1. Srinivas, K.," *Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques*", IEEE Transaction on Computer Science and Education (ICCSE),2010; (1344 -1349).
2. Yanwei Xing, "*Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease*", IEEE Transactions on Convergence Information Technology, Nov. 2007; (868 –872): 21-23
3. IBM, "*Data mining techniques*", http://www.ibm.com/developerworks/opensource/library/ba-data mining techniques /index.html?ca=drs-, downloaded on 04 April 2013.
4. Glymour C., D. Madigan, D. Pregidon andP.Smyth, "*Statistical inference and data mining*",Communication of the ACM, 2006; 35-41.
5. Thuraisingham, B.:"*A Primer for Understanding and Applying Data Mining*", IT Professional, 2000; 28-31.
6. Han, J., Kamber, M.: "*Data Mining Concepts and Techniques*", Morgan Kaufmann Publishers, 2006.
7. Ho, T. J.: "*Data Mining and Data Warehousing*", Prentice Hall, 2005.
8. Sriram. N: " *A Study on Machine Learning Techniques for Data Mining*" International Journal of Basic and Applied Research , June 2008.

9. Srinivas,K., "*Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques*", IEEE Transaction on Computer Science and Education (ICCSE), 2010; (1344 -1349), .
10. Shantakumar, B.Patil,Y.S.Kumaraswamy,"*Predictive data mining for medical diagnosis of heart disease prediction*",IJCSE 2011; 17.
11. M. Anbarasi et. al. "*Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm*",International Journal of Engineering Science and Technology 2013; 2(10): 5370-5376 .
12. HninWintKhaing, "*Data Mining based Fragmentation and Prediction of Medical Data*", IEEE,2011.
13. V.Chauraisa and S. Pal, "*Data Mining Approach to Detect Heart Diseases*", International Journal of Advanced Computer Science and Information Technology (IJACSIT),2013; 2(4): 56-66.
14. S. Halder and A. Ghosal, "*A survey on mobility-assisted localizationtechniques in wirelesssensor networks*," J. Netw. Comput. Appl., Jan 2016; 60:82–94.
15. Quinlan J. "*Induction of decision trees"*. Mach Learn 1986; 1:81—106.
16. Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. "*Classification and regression trees. Monterey*", CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.
18. AnandBahety,"*Extension and Evaluation of ID3 –Decision Tree Algorithm*". University of Maryland, College Park.
19. S. K. Yadav and Pal S., "*Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*", World of Computer Science and Information Technology (WCSIT), 2012; 2(2): 51-56.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. "*Classification and regression trees*". Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.
21. AnandBahety, "*Extension and Evaluation of ID3 –Decision Tree Algorithm*". University of Maryland, College Park.
22. Dangare, C. S., &Apte, S. S. "*Improved study of heart disease prediction system using data mining classification techniques.International Journal of Computer Applications"*,2012; 47(10): 44-48.
23. Nagalakshmi, D., &Balayesu, N.. "*Predicting Heart Disease Using Datamining Techniques".Ijitr*,2017; 5(4): 7036-7037.
24. Patel, A., Gandhi, S., Shetty, S., &Tekwani, B. "*Heart Disease Prediction Using Data*

*Mining"*2017.

25. Sultana, M., Haider, A., &Uddin, M. S. (2016, September) "*Analysis of data mining techniques for heart disease prediction*".In Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on:1-5. IEEE.