International Journal of Recent Development in Computer Technology & Software Applications
Vol. 1, Issue 1-2017

DATA MINING FOR HEART DISEASE PREDICTION SYSTEM

MS MHASKE*, PS BHUSARI*, PS INGLE*

ABSTRACT

Data mining which is also known as knowledge discovery is the process in which we extract useful information from the large set of the data. Use data mining classification modeling technique, namely, decision trees. In data mining technique the decision tree approach is more powerful for classification problems. A decision tree made of a root node, branches and leaf nodes. To evaluate the data, follow the path from root node to reach leaf node. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48.This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Heart disease leading cause of death in the world over the past 10 years, about 25% of deaths in the age group of 25-69 years occur because of heart disease. Heart disease is major disease all over world. In medical science prediction of heart disease is very important. Applying decision tree technique to heartdiseaseprediction data can provide as reliable performance to achieve in diagnosing heart disease. In data mining, decision tree techniques are used to find heart disease of patient. Based on the risk factors the heart disease can predict.

KEYWORDS: Data Mining, Classification, Decision Tree, ID3, C4.5, CHAID, C5.

INTRODUCTION

Heart disease is a major cause of morbidity and mortality. According to the World Health Organization, 12 million deaths are caused by heart diseases in the world annually, 50 percent of which can be prevented by controlling risk factors. Heart diseases are expected to be the main reason for 35 to 60 percent of total deaths expected worldwide by 2025. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate.Today diagnosing patients correctly and administering effective treatments have become quite a challenge. Poor clinical decisions may end to patient death and which cannot be afforded by the hospital as it loses its reputation.

*Department of Information Technology, Anuradha Engineering College, Chikhli.
Correspondence E-mail Id: editor@eurekajournals.com

The cost to treat a patient with a heart issue is very high and not reasonable by each patient. To accomplish a right and practical treatment PC based data as well as choice emotionally supportive networks can be produced to do the assignment. Most clinics today utilize a type of doctor's facility data frameworks to deal with their medicinal services or patient information. Because of persistent expanding the extent of medicinal services information a sort of many-sided quality exist in it. At the end of the day, we can state that medicinal services information turns out to be extremely mind boggling. By utilizing the conventional techniques it turns out to be extremely troublesome to extricate the significant data from it. Information mining is the way toward extricating concealed learning from information. It can uncover the examples and connections among huge measure of information in a solitary or a few datasets. Information mining innovation gives a client situated way to deal with novel and concealed examples in the information.

Data mining is the nontrivial process of identifying valid novel potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining is also known as knowledge discovery process consist of an iterative sequence of data cleaning ,data integration, data selection, data mining, pattern recognition and knowledge presentation. In the Fast moving world people want to live a very luxuries life so they work like a machine in order to earn Lot of money and live comfortable life therefore in this technique race they forget to take care of themselves.In the features that increase the possibility of heart attack are smoking, lack of physical exercises, highblood pressure, high cholesterol, and healthy diet, harmful use of alcohol and high sugar levels. Prediction in data mining involves attributeor variable in the data set to find anunknownor future state algorithm. The current research intends to predict the probability of getting heart disease given patient data set. Disease prediction plays an important role in data mining. Effective and efficient automated heart disease prediction system can be beneficial in healthcare sector for heart disease prediction. In this analyzes the heart disease prediction using decision tree technique.

For detecting a disease number of test should be required from the patient but with the help of data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden pattern and also provided healthcare professionals an additional source of knowledge for making decisions. Healthcare information system is being used in almost all of the hospitals so as to manage health care data; as the system consists of enormous amount of data used to extract concealed data to building intelligent medical diagnosis. The pronominal aim is to develop prediction system which will diagnose heart disease by making use of patient’s medical dataset.

LITERATURE SURVEY

- In 2008 Sellappan Palaniappan, Rafiah Awang, An Intelligent Heart Disease Prediction System (IHDPS) is created by utilizing information mining methods Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al [1]
- In 2009 Sitar-Taut, V.A. et al Naïve Bayes, Decision tree is an information mining strategy that shows achievement in order of diagnosing coronary illness patients[2]
- In 2010 Bharati M. Ramageri, The classifier-preparing calculation utilizes these pre-arranged cases to decide the arrangement

of parameters required for legitimate segregation. The calculation at that point encodes these parameters into a model called a classifier [3]

- In 2011 K. Srinivas et al. introduced Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks. The potential utilization of grouping based information mining methods, for example, Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to the huge Volume of human services data.[4]
- In 2012 Sudha et al. to propose the order calculation like Naïve Bayes, Decision tree and Neural Network for foreseeing the stroke sicknesses. The grouping calculation like choice trees, Bayesian classifier and back spread neural system were embraced in this study.[5]
- In 2012 Chaitrali S. Dangare et.al broke down expectation framework for Heart ailment utilizing more number of qualities. This paper included two more quality stoutness and smoking. They communicated various elements that expansion the danger of Heart infection. That are , High Blood Cholesterol, Smoking, Family History, Poor Diet, Hyper Tension ,High Blood Pressure, Obesity and Physical latency. The information mining arrangement systems called Decision Tree, Naïve Bayes and Neural Network are broke down on Heart Disease database. The exhibitions of these procedures are analyzed in light of their precision. They utilized J48 calculation for this framework. J48 calculation utilizes pruning technique to fabricated a tree. This procedure gives most extreme precision on preparing data.[6]
- In 2013 Shamsher Bahadur Patel, Pramod Kumar Yadav, and Dr. D. P. Shukla has displayed an examination paper "Anticipate the Diagnosis of Heart Disease Patients Using Class1ification Mining Techniques". In this exploration paper, the social insurance industry, the information digging is for the most part utilized for the forecast of coronary illness. The target of our attempts to anticipate the finding of coronary illness with a diminished number of traits utilizing Decision Tree.[7]
- In 2013Atul Kumar Pandey et al. proposed a forecast demonstrate with 14 traits. they built up that model utilizing j48 Decision Tree for ordering Heart Disease in view of the Clinical highlights against unpruned, pruned and pruned with decreased mistake pruning method.[8]
- In 2015 MoloudAdbar, Sharareh R. NiakanKalhori, ToleSutikno, Imam Much IbnuSubroto, goliArji exhibited the exploratory outcomes and utilized different information mining methods like C5.0, Neural Network, Support Vector Machine, KNN.[9]

DATA MINING

Data Mining is core part of Knowledge Discovery Database (KDD). Numerous individuals regard Data Mining as an equivalent word for KDD since it's a key piece of KDD process. Information disclosure as a procedure is delineated in Figure 1 and comprises of an iterative grouping of the accompanying advances:

- DATA CLEANING: To expel clamor or unessential information.
- DATA INTEGRATION: Where numerous information sources might be joined.
- DATA SELECTION: Where information important to the investigation undertaking are recovered from the database.
- DATA TRANSFORMATION: Where information are changed or solidified into shapes fitting for mining by performing synopsis / total activities.
- DATA MINING: A fundamental procedure where canny techniques are connected

keeping in mind the end goal to extricate information Patterns.

- PATTERN EVALUATION: To recognize the really fascinating examples speaking to learning in light of some intriguing quality measures.

KNOWLEDGE PRESENTATION

Knowledge representation techniques are used to present the minedknowledge to the user.[10]

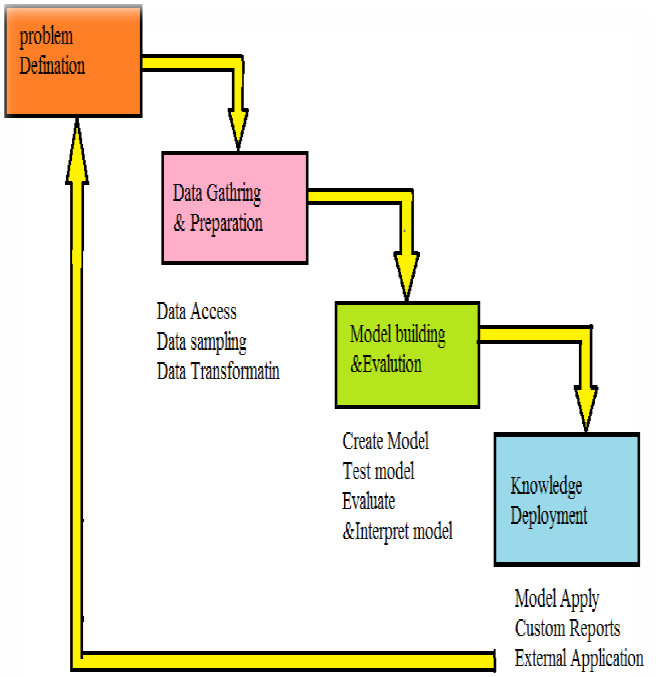

Figure 1.Data Mining Process

HEART DISEASE

The heart is vital organ or part of our body. Life is itself subject to productive working of heart. operation of heart is not proper, it will influence the other body parts of human, for example, mind, kidney and so forth. It is simply a pump, which directs blood through the body. In the event that flow of blood in body is wasteful the organs like cerebrum endure and if heart Life is totally reliant on effective working of the heart. The term Heart sickness alludes to infection of heart and vein framework inside it.

There are number of factors which increase the risk of Heart disease

- Family history of heart disease
- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

SYMPTOMS OF A HEART ATTACKCAN INCLUDE

- Discomfort, weight, greatness, or agony in the chest, arm, or beneath the breastbone.
- Discomfort emanating to the back, jaw, throat, or arm.
- Fullness, acid reflux, or stifling feeling (may feel like indigestion).
- Sweating, sickness, heaving, or wooziness.

- Extreme shortcoming, tension, or shortness of breath.
- Rapid or unpredictable heartbeats Types[11]

METHEDOLOGY

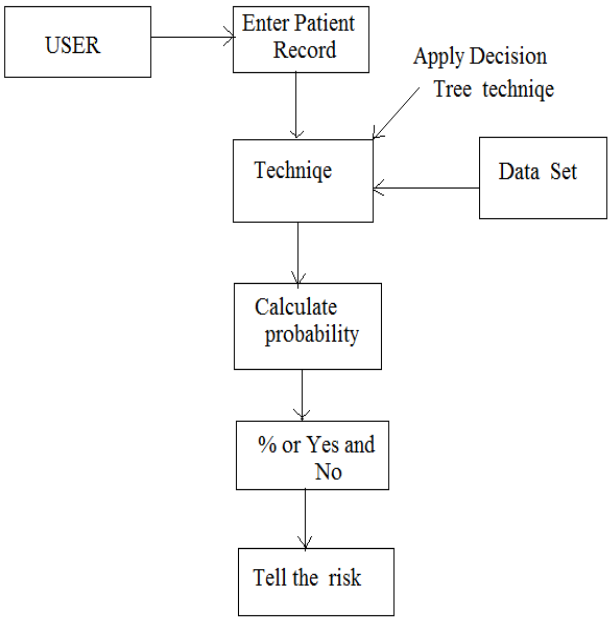

Figure 2.Heart Disease Prediction System

- From fig. Heart Disease Prediction System User can take the input as patient data these data give from doctor or medical institute.
- First enter the record of patient then applying the data mining technique such as Decision tree inpatient data.
- Applying the technique compare the data set and available from the medical institute or doctors.
- Compare the each attribute with previous or historical dataset. The data set must be trained using Software such as weka.
- Then check the symptoms of patient in the patient data set. In the data source symptoms of patient is given as follow:

DATA SOURCE

1. Age: It will take age in years as input.
2. Sex: It will take two values as input i.e.
   - Value 1: Male and
   - Value 0: Female
3. Chest Pain Type: It will take four value as input which shows the chest pain type as
   - Value 1: typical type-1 angina,
   - Value 2: typical type angina,
   - Value 3: non-angina pain;
   - Value 4: asymptomatic.
4. Trest Blood Pressure: resting blood pressure (in mm Hg on admission to the hospital)
5. Chol: serum cholesterol in mg/dl.
6. Fasting Blood Sugar: It will take two values as input i.e.
   - Value 1 for FBS > 120 mg/dl and
   - Value 0:for FBS < 120 mg/dl.
7. Restecg: Resting electrographic results will take 3 value as input i.e.
   - Value 0: normal;
   - Value 1: 1 having ST-T wave abnormality;
   - Value 2: showing probable or definite left ventricular hypertrophy.

8. Thalach: Maximum heart rate achieved by the patient.

9. Exang: Exercise induced angina will take two values as input i.e.
   - Value 1: yes and
   - Value 0: no.

10. Old peak: ST depression induced by exercise relative to rest.

11. Slope: The slope of the peak exercise ST segment will take three values as input i.e.
    - Value 1: unsloping.
    - Value 2: flat and
    - Value 3: down sloping.

12. CA: Number of major vessels colored by fluoroscopy will take three values as input (value 0-3)

13. Thai: It will take 3 input values i.e.
    - Value 3: normal
    - Value 6: fixed defect and
    - Value 7: reversible defect.

14. Num: This a diagnosing attribute having two input value i.e Value 0: < 50% diameter narrowing (no heart disease); and
    - Value 1: > 50% diameter narrowing (has heart disease).[12]

- While comparing the patient data to the available data set tells user about the symptoms ofpatient

- Later by using Data mining technique we get probabilityin the form of yes or no and thepatient having heart attack or not.
- There is no risk in this technique.
- We just check the symptoms of heart disease.
- If the patient found the symptoms of heart disease.
- You will take a treatment from the doctor.
- Eg. In the data mining technique Patient record enter in the system the patient record decision tree how work. Enter Age of patientin the data set and other attribute and check the heart disease.

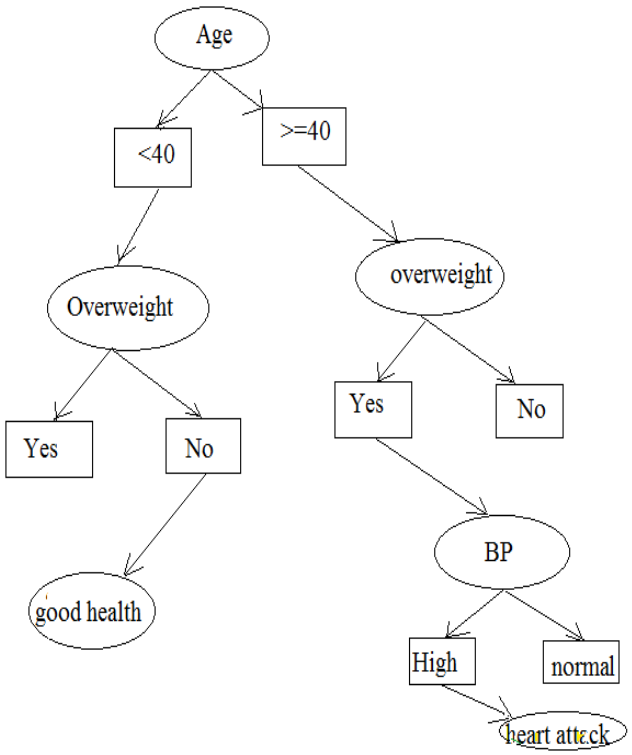

Figure 3.Decision Tree Example

DATA MINING TECHNIQE

CLASSIFICATION

Classification data mining technique based on machine learning. Used to classify each item in a set of data into one of predefined set of classes or groups. Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency.[13]

DECISION TREE

Berry and Linoff characterized decision tree as "a structure that can be utilized to separate up a vast accumulation of records into progressive littler arrangements of records by applying a grouping of basic choice principles. With each progressive division, the individuals from the subsequent sets turn out to be increasingly like each other." The decision tree approach is all the more effective for characterization issues. There are two stages in this procedures constructing a tree and applying the tree to the dataset. There are numerous well known decision tree calculations CART, ID3, C4.5, CHAID, and J48.

Decision tree is like the flowchart in which each non-leaf hubs signifies a test on a specific property and each branch indicates a result of that test and each leaf hub have a class name. The hub at the best most names in the tree is called root hub. Utilizing Decision Tree, chiefs can pick best option and traversal from root to leaf demonstrates remarkable class detachment in light of greatest data pick up.

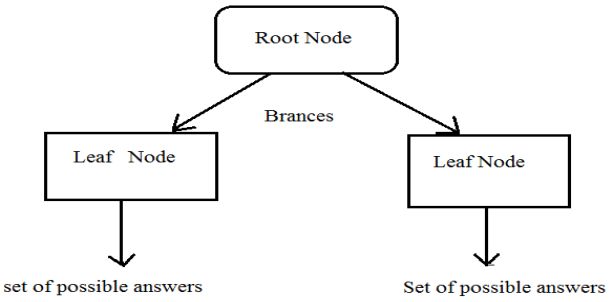

Figure 4.Decision Tree

DECISION TREE ALGORITHM

ID3(ITERATIVE DICHOTOMISER)

ID3 the word stands for Iterative Dichotomiser 3. ID3 is one of the decision tree models that build a decision tree from a fixed set of training instances. The resulting tree is used to classify the future samples. Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. The main advantage of this algorithm it that accepts only categorical attributes and only one attribute tested at a time for making decision.

C4.5

C4.5 is a calculation used to produce a choice tree created by Ross Quinlan the choice trees created by C4.5 can be utilized for grouping and hence C4.5 is regularly alluded to as a measurable classifier. With the goal that C4.5 is regularly called as Statistical Classifier. C4.5 is an augmentation of Quinlan's prior ID3 calculation. C4.5 is the most recent variant of ID3 calculation. This constructs a choice tree like the ID3. It can acknowledge information with straight out or numerical esteems. C4.5

calculation can without much of a stretch handle missing esteems, as missing quality esteems are not used by C4.5.

C5.0

This model is a change of C4.5 choice tree calculation. Both C4.5 and C5.0 can create classifiers communicated as either choice tree or run sets. In numerous applications, control set are favored on the grounds that they are less complex and less demanding to get it. The real contrasts are tree sizes and calculation time. C5.0 is utilized to deliver littler trees and quick than C4.5.It produces basic and little choice trees.

J48

J48 choice tree is the execution of ID3 calculation created by WEKA venture group. J48 is a basic C4.5 choice tree for characterization. With this strategy, a tree is built to demonstrate the grouping procedure. Once the tree is manufacture, it is connected to each tuple in the database and the outcome in the order for that tuple. In the WEKA information mining device, J48 is an open source Java usage of the C4.5 calculation. In different calculations the arrangement is performed recursively till each and every leaf is unadulterated, that is the grouping of the information ought to be as flawless as could be expected under the circumstances. This calculation it creates the tenets from which specific character of that information is produced. The goal is logically speculation of a choice tree give adaptability and precision.

TRUCK

It remains for Classification and Regression Trees. It was presented by Breiman in 1984. It manufactures the two characterizations and relapse trees. The characterization tree development via CART depends on parallel part of the traits. The relapse investigation highlight is utilized as a part of guaging a needy variable given an arrangement of indicator factors over a given timeframe. Trucks underpin nonstop and ostensible characteristic information and have normal speed of preparing.

CHAID (CHI-SQUARED AUTOMATIC INTERACTION DETECTION)

CHAD is a sort of choice tree system, in view of balanced essentialness testing the method was created in south Africa and was distributed in 1980 by Gordon V. Kass CHAID in view of a formal expansion of the US AID (Automatic Interaction Detection) and THAID (THeta Automatic Interaction Detection) strategies of the 1970s, which thusly were augmentations of prior research, incorporating that performed in the UK in the 1950s. CHAID is frequently utilized as a part of the setting of direct advertising to choose gatherings of purchasers and anticipate how their reactions to a few factors influence different factors, albeit other early applications were in the field of restorative and mental research.

Like other choice trees, CHAID's points of interest are that its yield is exceptionally visual and simple to decipher. Since it utilizes multiway parts of course, it needs rather substantial example sizes to work successfully, since with little example sizes the respondent gatherings can rapidly turn out to be too little for dependable investigation.[14]

OPEN SOURCE TOOLS FOR DATA MINING

WEKA TOOL

WEKA is a data mining framework created by the University of Waikato in New Zealand that executes information mining calculations utilizing the JAVA dialect. WEKA is a cutting edge office for creating machine learning systems and their application to certifiable data mining issues. It is an accumulation of machine

learning calculations for data mining errands. The calculations are connected specifically to a dataset. WEKA executes calculations for data preprocessing, grouping, relapse, and bunching and affiliation rules. It additionally incorporates perception devices. The new machine learning plans can likewise be produced with this bundle. WEKA is open source programming issued.

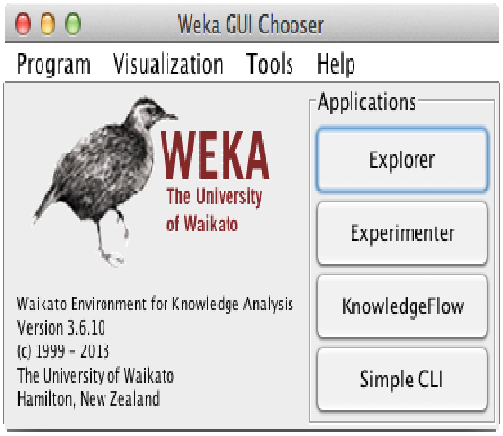

Figure 5.Weka Tool

TANAGRA

Tanagra is free data mining software for scholarly and inquires about purposes. It proposes a few data mining strategy from exploratory data examination, accurate learning, machine learning and database zone.Tanagra is an open source venture as each specialist can access to the source code and include his own particular calculations, to the extent he concurs and fits in with the product circulation permit. The primary motivation behind Tanagra venture is to give analysts and understudies a simple to utilize data mining programming, fitting in with the present standards of the product improvement in this area and permitting to dissect either genuine or engineered information.

ORANGE

Orange is an open data perception and investigation for beginner and specialists. Data mining utilized through visual Programming python on scripting, segment for machine learning. Include ones utilized for bioinformatics and content mining. This is pressed with highlights for data examination.

.NET FRAMWORK

.net structure is a product system created by Microsoft that runs basically on Microsoft windows and gives dialects interoperability over a few programming dialects. For engineers the .NET Framework gives an exhaustive and reliable application that has outwardly dazzling client encounters and consistent and secure correspondence.

ADVANTAGES OF HEART DISEASE PREDICTION SYSTEM

- Reduce the fraud while checking The Disease.
- It can provide the probability of Disease.
- The technique gives more accuracy then the chance of disease is high.
- If the data is high then it gives more accuracy.
- Reduce problem complexity.
- Chip in cost
- Easy to uses.
- Reduced Time.
- They are simple to understand and interpret. People are able to understand decision tree models after a brief explanation

DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

Medicinal services industry today creates a lot of complex information about patients, healing center assets, malady analysis, electronic patient records, and restorative gadgets and so on. Bigger measures of information are a key asset to be handled and examined for learning extraction that empowers bolster for cost-reserve funds and basic leadership. Information mining applications in social insurance can be gathered as the assessment into general classes.

TREATMENT EFFECTIVENESS

Information mining applications can create to assess the adequacy of therapeutic medicines. Information mining can convey an examination of which strategy demonstrates compelling by looking into causes, side effects, and courses of medicines.

HUMAN SERVICES MANAGEMENT

Information mining applications can be created to better distinguish and track incessant ailment states and high-chance patients, outline proper mediations, and lessen the quantity of clinic confirmations and cases to help social insurance administration. Information mining used to break down gigantic volumes of information and insights to look for designs that may show an assault by bio-psychological oppressors.

CLIENT RELATIONSHIP MANAGEMENT

Client relationship administration is a center way to deal with overseeing communications between business association ordinarily banks and retailers-and their clients, it is no less critical in a social insurance setting. Client cooperation’s may happen through call focuses, doctors' workplaces, charging offices, inpatient settings, and wandering consideration settings.

EXTORTION AND ABUSE

Recognize misrepresentation and misuse set up standards and afterward distinguish unordinary or irregular examples of cases by doctors, facilities, or others endeavor in information mining applications. Information mining applications extortion and manhandle applications can feature wrong medicines or referrals and deceitful protection and restorative cases.[16]

CONCLUSION

This seminar concludes that Heart Disease is a fatal disease by its nature, and leading cause of death for men and women. This disease makes a life threatening complexities such as heart attack and death. Know the warning signs and symptoms of heart attack. The chances of survival are greater when treatment begins quickly. Importance of Data Mining in the Medical Domain is realized and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied. The observations from the previous work have led to the deployment of the proposed system architecture for this work. Through, various classification techniques are widely used for Disease Prediction, Decision Tree classifier is selected for its simplicity and accuracy

REFRENCES

[1]. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System UsingData Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[2]. Sitar-Taut, V.A., et al., "Using machine learning algorithms in cardiovascular disease risk evaluation”, Journal of Applied Computer Science, 2009.

[3]. Bharati M. Ramageri, "Data Mining Techniques and Applications": Indian Journal ofcomputer Science and Engineering Vol. 1 No. 4 301-305, 2010.

[4]. K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Application of Data Mining Techniques inHealthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp. 250-255, 2011.

[5]. A. Sudha, P. Gayathiri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", International Journal of Computer Applications, Vol. 43, No. 14, pp. 0975-8887, 2012.

[6]. Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction Using Data Mining Classification Techniques", International Journal of Computer Applications, Vol. 47,No. 10, pp. 0975-888, 2012.

[7]. ShamsherBahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" IOSR Journal of Agriculture and Veterinary Science (IOSRJAVS),Volume 4, Issue 2 (Jul.-Aug. 2013).

[8]. Atul Kumar Pandey, PrabhatPandey, K.L. Jaiswal and Ashok Kumar Sen, " A Heart DiseasePrediction Modelusing Decision Tree", IOSR Journal of Computer Engineering, Vol. 12, Issue.6, (Jul.-Aug. 2013), pp. 83-86.

[9]. MoloudAdbar, Sharareh R. NiakanKalhori, toleSutikno, Imam Much IbnuSubroto, GoliArji "Comparing Performance of DataMining algorithms in Prediction Heart Diseases", IJECE, Vol 5,No 6, December 2015.

[10]. Beant Kaur, Williamjeet Singh "Review on Heart Disease Prediction System using DataMining Techniques" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 10 IJRITCC October 2014, ISSN: 2321-8169www.ijritc.org.

[11]. K.Sudhakar, Dr. M. Manimekalai "Study of Heart Disease Prediction using Data MiningInternational Journal of Advanced Research in Computer Science and Software EngineeringResearch Paper Volume4, Issue1 ,January2014 Available online at: www.ijarcsse.comAvailable online at: www.ijarcsse.com.

[12]. R. Vijaya Kumar Reddy, K. PrudviRaju "Prediction of Heart Disease Using Decision Tree Approach" International Journal of Advanced Research in Computer Science and Software Engineering Research paper, Volume 6, Issue 3, March 2016, ISSN: 2277 128X Available online at: www.ijarcsse. com.

[13]. K. Thenmozhi , P. Deepika "Heart Disease Prediction Using Classification with DifferentDecision Tree Techniques" International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014ISSN 2091-2738 www.ijergs.org.

[14]. SonamNikhar1, A. M. Karandikar "Prediction of Heart Disease Using Data Mining Techniques"-A Review International Research Journal of Engineering and Technology (IRJET), Volume: 03Issue: 02Feb-2016, e-ISSN: 2395-0056 www.irjet.net.

[15]. Himani Sharma, Sunil Kumar "A Survey on Decision Tree Algorithms of Classification inData Mining"International Journal of Science and Research, Volume 5,Issue 4,April 2016 www.ijsr.net , (IJSR) ISSN (Online): 2319-7064.

[16]. M Durairaj, V. Ranjani "Data Mining Applications In Healthcare Sector: A Study"International Journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013, ISSN 2277-861629, www.ijstr.org.

