A Proficient Heart Disease Prediction Method Using Different Data Mining Tools
K. Manimekalai
Head and Assistant Professor
Department of BCA
Sri GVG Visalakshi College for Women, Udumalpet, Tamilnadu, India.
gvgmanimekalai@gmail.com

Abstract:
Heart disease is a major health problem and it affects a large number of people. In medical, prediction of heart disease is very important. In order to save a patient's life, lot of effort is being taken by the hospitals and medical practitioners. The health sector today contains hidden information that can be important in making decisions. Diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. Data mining techniques are used to analyze this rich collection of data from different perspectives and deriving useful information. The objective of this paper is to study the various data mining tools available to predict the heart disease and find the best method of prediction.

Keywords: Data Mining, .NET data mining tool, TANGARA, WEKA, Decision Trees, Fuzzy Logic, J48 Technique, Naive Bayes, Neural Network.

I. Introduction
Heart attack diseases remain the main cause of death worldwide. Medical practitioners generate data with a wealth of hidden information present, and it’s not properly being used effectively for predictions. For this purpose, the research converts the unused data into a dataset for modeling using different data mining techniques. People die having experienced symptoms that were not taken into considerations. There is a need for medical practitioners to predict heart disease before they occur in their patients. The features that increase the possibility of heart attacks are smoking, lack of physical exercises, high blood pressure, high cholesterol, unhealthy diet, harmful use of alcohol, and high sugar levels. Cardio Vascular Disease (CVD) incorporates coronary heart, cerebrovascular (Stroke), hypertensive heart, congenital heart, peripheral artery, rheumatic heart, inflammatory heart disease.

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.

Identifying the major risk factors of Heart Disease categorizing the risk factors in an order which causes damages to the heart such as high blood cholesterol, diabetes, smoking, poor diet, obesity, hyper tension, stress, etc. Data mining functions and techniques are used to identify the level of risk factors to help the patients in taking precautions in advance to save their life.[18]

II Literature Survey
Numerous studies have been done that have focused the prediction of heart disease. They have used different data mining tools and applied different techniques for diagnosis and achieved different results.

In the year 2013, Abhishek et al used the data mining tool Weka 3.6.4 in heart disease prediction system. They used J48 technique and achieved 95.56% accuracy and using Naive Bayes achieved 92.42%. [10]

In 2013 Rashedur et al used Neural network technique using Weka data mining tool and achieved 79.19% and to compare various classification techniques. They used another technique fuzzy logic using TANGRA data mining tool and achieved 83.85% accuracy. [11]

In the year 2012, Nidhi et al used data mining tool Weka 3.6.6 in the analysis of heart disease prediction system and achieved 99.52% using Naive Bayes. She also used TANGRA data mining tool but could achieve up to 52.33% only using decision trees. They used .NET data mining tool and achieved up to 96.5% using neural networks. [14]

In the year 2008, Humar et al used classification, Backpropagation, Fuzzy neural network techniques for diabetes and heart diseases. [15]

Hlaudi Daniel Masethe et al[22] used the Data Mining algorithm such as J48, Naive Bayes, REPTREE, CART, and Bayes Net are applied in this research for predicting heart attacks. The research result shows prediction accuracy of 99%.
Ms. Rupali R. Patil [23] explained their paper that Decision Support in Heart Disease Prediction System is developed using both Naive Bayesian Classification and Jelinek-mercer smoothing technique. The system extracts hidden knowledge from a historical heart disease database. Jelinek-mercer smoothing technique is the more effective than naive bayes to predict patients with heart disease.

In year 2010, a study was conducted for predictive model for the Ischemic Heart Disease (IHD); they applied Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN) and the support vector machine (SVM) to develop classification models for identifying IHD patients on a data obtained from measurements of cardiac magnetic field at 36 locations (6 x 6 matrices) above the torso. The result shows that BPNN and BNN gave the highest classification accuracy of 78.43 %, while RBF kernel SVM gave the lowest classification accuracy of 60.78 %. BNN presented the best sensitivity of 96.55 % and RBF kernel SVM displayed the lowest sensitivity of 41.38 %. Both polynomial kernel SVM and RBF kernel SVM presented the minimum and maximum specificity of 45.45 % and 86.36 %, respectively.

Abhishek Taneja [24], designed a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography.

B. Venkatalakshmi [25] implemented the Data set of 294 records with 13 attributes is used and the outcome reveals that the Naive Bayes outperforms and sometime Decision Tree.

K. Manimekalai [19] provide a study of different input attributes that can be used to predict the heart disease. Some attributes are mainly responsible for heart disease. In order to predict the disease based upon patient’s raw symptom description, different input parameters can be collected. To help the physicians and healthcare professionals in the prediction of heart diseases, the number of attributes should standard. The system helps the medical practitioners to minimize the time of prediction and improve their practice as well as protect the life of the patient.

K. Manimekalai [20] provide a study of Heart Diseases using various Data Mining Techniques. When the data mining technique is used separately the accuracy is low. To improve the accuracy value, data mining techniques should be combined together. SVM Classifier with genetic Algorithm contains 95% high accuracy than other technique. In future the techniques are hybrid, the accuracy will high. It will definitely help the patients as well as the medical practitioners to predict the heart disease.

III. Proposed Prediction System

In this paper different versions of Weka Tool, TANGARA and .NET data mining tools are discussed. Abhishek et al used Weka 3.6.4 data mining tool. They used J48 Techniques is used. They achieved 95.56% result. They used another technique Naive Bayes and achieved 92.42%.

IV Data Mining Tools

There are various data mining tools used for data mining purpose. These are WEKA, TANAGRA, MATLAB and .NET FRAMEWORK.

WEKA: It is a data mining tool which was developed in New Zealand by the University of Waikato that implements data mining algorithms using JAVA language. WEKA is a collection of machine learning algorithms and their application to the data mining problems. These algorithms are directly applied to the dataset. WEKA supports data file in ARFF format. WEKA is open source software and hence, it is not dependent on any platform. It includes algorithms for data processing, classification, regression, clustering, association and also visualization tools.

TANAGRA: It is open source software as researchers can access to the source code and add their own algorithms and compare their performances, if it conforms to the software distribution license. It includes several data mining algorithms from statistical learning, machine learning, data analysis and database area.

MATLAB: It is a data mining tool built in high level language. It provides interactive environment for visualization, numerical computation and programming. The built in math functions, language and tool explore various approaches and helps to reach a solution faster than with the spreadsheet of traditional programming languages like C, C++.

<p>| Table I: DIFFERENT DATA MINING TOOLS USED IN HEART DISEASE PREDICTION SYSTEM WITH ACCURACY |</p>
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Data Mining tool</th>
<th>Techniques Used</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Abhishek et al</td>
<td>Weka 3.6.4</td>
<td>J48 Technique</td>
<td>95.56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Naive Bayes</td>
<td>92.42%.</td>
</tr>
<tr>
<td>2013</td>
<td>Rashedur et al</td>
<td>Weka</td>
<td>Neural Network</td>
<td>79.19%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TANAGRA</td>
<td>83.85%</td>
</tr>
<tr>
<td>2012</td>
<td>Nidhi et al</td>
<td>Weka 3.6.6</td>
<td>Naive Bayes</td>
<td>99.52%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TANAGRA</td>
<td>52.33%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.NET data mining tool</td>
<td>96.5%</td>
</tr>
</tbody>
</table>

In the year 2013, Rashedur et al used Weka tool and they used the technique Neural Network and achieved 79.19%. They tried another tool TANGARA and the technique is FUZZY Logic and they receive 83.85%. In the year 2012, Nidhi et al used three data mining tools namely Weka 3.6.6 and they used the technique is Naive Bayes and achieved the result 99.52%. The data mining tool TANGRA and they used the technique is Decision Tree and result is 52.33%. And another data mining tool is Decision Tree. And they used a new tool .NET data mining tool with Neural Network and the result is 96.5%.
and JAVA. It analyse data, develop algorithms, and create models and applications.

.NET FRAMEWORK: It is a software framework developed by Microsoft which runs primarily on Microsoft Windows. It provides secure communication and consistent applications. It provides language interoperability (each language can code written in other languages) across several programming languages.

V Data Mining Techniques

i) J48 Technique

Decision tree is a kind of classifying and predicting data mining technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods. Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis, predicting essential hypertension, and predicting cardiovascular disease. Decision tree is one of the most popular tools for classification and prediction. Production of a decision tree is an efficient method for classification of data. This tree using a top-down strategy to build a test on each node. J48 decision tree method is the implementation of c4.5 decision tree in weka data mining tool. J48 decision tree supports continuous and discrete features. It can also manage features with missing value.

ii) Naïve Bayes Algorithm

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. A naïve Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It learns from the “evidence” by calculating the correlation between the target (i.e., dent) and other (i.e., independent) variables. Depending on the precise nature of the probability model, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. Naïve Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. For example, a patient may bed to have certain symptoms. Based on the observation, Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct. Bayes’ Theorem finds the probability of an even occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes theorem can be stated as follows, P (B given A) = Prob(A and B) / Prob (A) To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur P (Ci/X)>P (Cj/X) for all 1< = j< = m and j!= i

iii) Neural Network

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron

iv) Fuzzy Logic

Crisp input values are transferred into fuzzy values in the stage of fuzzification. The Fuzzy values are taken into the Generating rules for Advanced Fuzzy Resolution Mechanism.

Table II . COMPARISION OF DATA MINING TOOLS WITH TECHNIQUES AND ACCURACY

<table>
<thead>
<tr>
<th>Data Mining tools</th>
<th>Techniques</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weka 3.6.4</td>
<td>J48 Technique</td>
<td>95.56%</td>
</tr>
<tr>
<td>Weka 3.6.4</td>
<td>Naïve Bayes</td>
<td>92.42%</td>
</tr>
<tr>
<td>Weka</td>
<td>Neural Network</td>
<td>79.19%</td>
</tr>
<tr>
<td>TANGRA</td>
<td>Fuzzy Logic</td>
<td>83.85%</td>
</tr>
<tr>
<td>Weka 3.6.6</td>
<td>Naïve Bayes</td>
<td>99.52%</td>
</tr>
<tr>
<td>TANGRA</td>
<td>Decision Trees</td>
<td>52.33%</td>
</tr>
<tr>
<td>.NET data mining tool</td>
<td>Neural Network</td>
<td>96.5%</td>
</tr>
</tbody>
</table>

From Table II, different versions of Weka tools, TANGRA, .Net data mining tools are discussed. When the techniques used are differ, the result will vary. When compare to other data mining tools Weka tool give the highest accuracy.

Chart I : COMPARISION OF DATA MINING TOOLS WITH TECHNIQUES

The Chart I is drawn for the different data mining tools with various techniques. When compare to other data mining tools, Weka 3.6.6 data mining tool is easy, and it will produce the highest accuracy. When the data mining Tool Weka 3.6.4 with J48Technique is used the accuracy is 95.56%. But at the same time, the technique Naïve Bayes is used the accuracy is 92.42%. The previous version of WEKA is used with Neural Network the accuracy is very low. i.e
79.19%. The Data Mining tool TANGARA with Fuzzy Logic is used the accuracy is 83.85%. Weka 3.6.6 with Naïve Bayes is used the accuracy is 99.52%. The data mining Tool TANGRA with the technique Decision Tree is used, the accuracy is very low, i.e 52.33% . Net data mining tool is used with Neural Network, the accuracy is 96.5%.

V Results and Discussion

The objective of this work is to provide a study of Heart Diseases with various Data Mining tools with various versions. When compare to other data mining tools, Weka 3.6.6 with Naïve Bayes Technique is simple, easy, and it will produce the highest accuracy. Weka Tool is very easy to implement. It will definitely help the medical practitioners to predict the heart disease. This tool will protect the life of the patient. It will also help the researcher to select the appropriate tool.

REFERENCES


18. Divya Kundra, Er. Navpreet Kaur, “Review on prediction system for heart diagnosis using Data Mining Techniques, IJLRET ISSN: 2454-50, ISSN: 2454-5031


**BIOGRAPHY OF AUTHOR**

Manimekalai K working as a Head & Assistant professor of Computer applications in Sri GVG Visalakshi College for Women, Udumalpet, Tamilnadu, India. She is interested in Medical datamining. She has 11 years of experience. She received her B.Sc in Physics degree from the University of Bharathiyar in 1998. She did her MCA in Bharathidasan University in 2003. She had completed M.Phil in Bharathidasan University in 2007. She has published a journal on “A Review on Prediction of Heart Diseases by Comparing Risk Factors in Data Mining”, IJCSIT, Feb 2016. She has also published a journal on “Prediction of Heart Diseases using Data Mining techniques”, IJIRCCE, Feb 2016.