"Intelligent Heart Disease Prediction System Using Data Mining Techniques"

*Ms. Ishtake S.H , ** Prof. Sanap S.A.

*Department of Copmputer science, MIT, Aurangabad, Maharashtra, India.
** Department of Computer science, MIT, Aurangabad, Maharashtra, India.
***Corresponding author: Mail:** ishtake_suvarna@rediffmail.com

Abstract: Today medical services have come a long way to treat patients with various diseases. Among the most lethal one is the heart disease problem which cannot be seen with a naked eye and comes instantly when its limitations are reached. 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. The cost to treat a patient with a heart problem is quite high and not affordable by every patient. To achieve a correct and cost effective treatment computer-based information and/or decision support Systems can be developed to do the task. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" So there is need of developing a master's project which will help practitioners predict the heart disease before it occurs.The diagnosis of diseases is a vital and intricate job in medicine. The recognition of heart disease from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. Thus the attempt to exploit knowledge and experience of several specialists and clinical screening data of patients composed in databases to assist the diagnosis procedure is regarded as a valuable option.

Keywords: Data mining, IHDPS, Decision Tree, Neural Network, Naive Bayes

1. INTRODUCTION:

The main objective of our project is to develop a prototype Intelligent Heart Disease Prediction System (IHDPS) using three data mining modelling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms.

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together [2]. Intelligent Heart Disease prediction System (IHDPS) using data mining techniques, namely, Decision

trees, Naïve Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable [3].

A wide variety of areas including marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining [11]. Numerous fields associated with medical services like prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data as well employ Data Mining methodologies [12]. Providing precious services at affordable costs is a major constraint encountered by the healthcare organizations (hospitals, medical centers). Valuable quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost [3]. Owing to the accessibility of integrated information through enormous patient repositories, there is a swing in the insight of clinicians, patients and payers from qualitative visualization of clinical data to demanding a finer quantitative analysis of information with the assistance of all supporting clinical and imaging data. For example; now the physicians can evaluate diagnostic information of a variety of patients with identical conditions. Similarly, they can as well verify their findings with the conformity of physicians working on similar cases from all over the world [6].. Medical history data comprises of a number of tests essential to diagnose a particular disease [13]. Clinical databases are elements of the domain where the procedure of data mining has develop into an inevitable aspect due to the gradual incline of medical and clinical research data. It is possible for the healthcare industries to gain advantage of Data mining by employing the same as an intelligent diagnostic tool. It is possible to acquire knowledge and information concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, data mining has developed into a vital domain in healthcare [14]. It is possible to predict the efficiency of medical treatments by building the data mining applications. Data mining can deliver an assessment of which courses of action prove effective [15] by comparing and evaluating causes, symptoms, and courses of treatments. The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [9], [16]. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [12]. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling

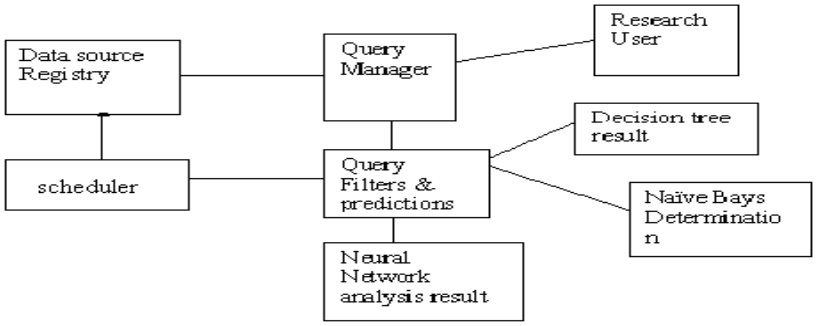


Figure 1. Block diagram of IHDPS

objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [6]. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [3]

2. **Mining methods** :

Decision Tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [7]. CART uses Gini index to measure the impurity of a partition or set of training tuples [6]. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods [14]. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. Neural Networks consists of three layers: input, hidden and output units (variables). Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. The higher the weight the more important it is. Neural Network algorithms use Linear and Sigmoid transfer functions. Neural Networks are suitable for training large amounts of data with few inputs. It is used when other techniques are unsatisfactory.

3. **Data source**: A total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database [1]. Figure 1 lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). To avoid bias, the records for each set were selected randomly. For the sake of consistency, only categorical attributes were used for all the three models. All the non-categorical medical attributes were transformed to categorical data. The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. The attribute "PatientID" was used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

Description of attributes :

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

4. **Validating model effectiveness**: The effectiveness of models was tested using two methods: Lift Chart and Classification Matrix. The purpose was to determine which model gave the highest percentage of correct predictions for diagnosing patients with a heart disease.

4.1 Lift Chart with predictable value. To determine if there was sufficient information to learn patterns in response to the predictable attribute, columns in the trained model were mapped to columns in the test dataset. The model, predictable column to chart against, and the state of the column to predict patients with heart disease (predict value = 1) were also selected. Figure 2 shows the Lift Chart output. The X-axis shows the percentage of the test dataset used to compare predictions while the Y-axis shows the percentage of values predicted to the specified state. The blue and green lines show the results for random-guess and ideal model respectively. The purple, yellow and red lines show the results of Neural Network, Naïve Bayes and Decision Tree models respectively. The top green line shows the ideal model; it captured 100% of the target population for patients with heart disease using 46% of the test dataset. The bottom blue line shows the random line which is always a 45-

degree line across the chart. It shows that if we randomly guess the result for each case, 50% of the target population would be captured using 50% of the test dataset. All three model lines (purple, yellow and red) fall between the random-guess and ideal model lines, showing that all three have sufficient information to learn patterns in response to the predictable state.

4.2 Lift Chart with no predictable value. The steps for producing Lift Chart are similar to the above except that the state of the predictable column is left blank. It does not include a line for the random-guess model. It tells how well each model fared at predicting the correct number of the predictable attribute. Figure 3 shows the Lift Chart output. The X-axis shows the percentage of test dataset used to compare predictions while the Y-axis shows the percentage of predictions that are correct. The blue, purple, green and red lines show the ideal, Neural Network, Naïve Bayes and Decision Trees models respectively. The chart shows the performance of the models across all possible states. The model ideal line (blue) is at 45-degree angle, showing that if 50% of the test dataset is processed, 50% of test dataset is predicted correctly

4.3 Classification Matrix.

Classification Matrix displaysthe frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. In this example, the test dataset contained 208 patients with heart disease and 246 patients without heart disease. Figure 4 shows the results of the Classification Matrix for all the three models. The rows represent predicted values while the columns represent actual values (1 for patients with heart disease, '0' for patients with no heart disease). The left-most columns show values predicted by the models. The diagonal values show correct predictions.

Counts for Decision Tree on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	219	62
1	27	146

Counts for Naive Bayes on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	211	28
1	35	180

Counts for Neural Network on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	211	30
1	35	178

Figure 3. Results of Classification Matrix for all the three models

5. Evaluation of Mining Goals: Five mining goals were defined based on exploration of the heart disease dataset and objectives of this research. They were evaluated against the trained models. Results show that all three models had achieved the stated goals, suggesting that they could be used to provide decision support to doctors for diagnosing patients and discovering medical factors associated with heart disease. The goals are as follows:

Goal 1: Given patients' medical profiles, predict those who are likely to be diagnosed with heart disease. All three models were able to answer this question using singleton query and batch or prediction join query. Both queries could predict on single input cases and multiple input cases respectively. IHDPS supports prediction using "what if" scenarios. Users enter values of medical attributes to diagnose patients with heart disease. For example, entering values Age = 70, CA = 2, Chest Pain Type = 4, Sex = M, Slope = 2 and Thal = 3 into the models, would produce the output in Figure 6. All three models showed that this patient has a heart disease. Naïve Bayes gives the highest probability (95%) with 432 supporting cases, followed closely by Decision Tree (94.93%) with 106 supporting cases and Neural Network (93.54%) with 298 supporting cases. As these values are high, doctors could recommend that the patient should undergo further heart examination. Thus performing "what if" scenarios can help prevent a

a potential heart attack.

Goal 2: Identify the significant influences and relationships in the medical inputs associated with the predictable state – heart disease. The Dependency viewer in Decision Trees and Naïve Bayes models shows the results from the most significant to the least significant (weakest) medical predictors. The viewer is especially useful when there are many predictable attributes. Figures 7 and 8 show that in both models, the most significant factor influencing heart disease is "Chest Pain Type". Other significant factors include Thal, CA and Exang. Decision Trees model shows 'Trest Blood Pressure" as the weakest factor while Naïve Bayes model shows 'Fasting Blood Sugar' as the weakest factor. Naïve Bayes appears to fare better than Decision Trees as it shows the significance of all input attributes. Doctors can use this information to further analyze the strengths and weaknesses of the medical attributes associated with heart disease.

Goal 3: Identify the impact and relationship between the medical attributes in relation to the predictable state – heart disease. Identifying the impact and relationship between the medical attributes in relation to heart disease is only found in Decision Trees viewer (Figure 9). It gives a high probability (99.61%) that patients with heart disease are found in the relationship between the attributes (nodes): "Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036." Doctors can use this information to perform medical screening on these four attributes instead of on all attributes on patients who are likely to be diagnosed with heart disease. This will reduce medical expenses, administrative costs, and diagnosis time. Information on least impact (5.88%) is found in the relationship between the attributes: "Chest Pain Type not = 4 and Sex = F". has the highest impact (92.58%). The least impact (0.2%) is found in the attributes: "Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >= 146.362 and < 158.036". Additional information such as identifying patients' medical profiles based selected nodes can also be obtained by using the drill through function. Doctors can use the Decision Tree viewer to perform further analysis.

Goal 4: Identify characteristics of patients with heart disease. Only Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Figure 10 shows that 80% of the heart disease patients are males (Sex = 1) of which 43% are between ages 56 and 63. Other significant characteristics are: high probability in fasting blood sugar with less than 120 mg/dl reading, chest pain type is asymptomatic, slope of peak exercise is flat, etc.Figure 11 shows the characteristics of patients with no heart disease with high probability in fasting blood sugar with less than 120 mg/dl reading, no exercise induced, number of major vessels is zero, etc. These results can be further analyzed.

Goal 5: Determine the attribute values that differentiate nodes favoring and disfavoring the predictable states: (1) patients with heart disease (2) patients with no heart disease. This query can be answered by analyzing the results of attribute discrimination viewer of Naïve Bayes and Neural Network models. The viewer provides information on the impact of all attribute values that relate to the predictable state. Naïve Bayes model (Figure 12) shows the most important attribute favoring patients with heart disease: "Chest Pain Type = 4" with 158 cases and 56 patients with no heart disease. The input attributes "Thal = 7" with 123 (75.00%) patients, "Exang = 1" with

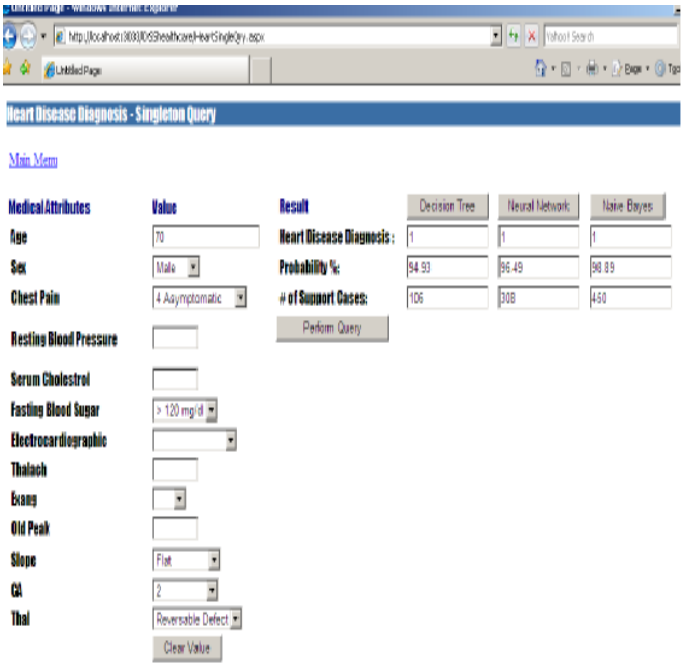
with 112 (73.68%) patients," Slope =2" with 138 (66.34%) patients, etc. also favor predictable state. In contrast, the attributes "Thal = 3" with 195 (73.86%) patients, "CA = 0" with 198 (73.06%) patients, "Exang = 0" with 206 (67.98%), etc. favor predictable state for patients with no heart disease.

6. Benefits and limitations

IHDPS can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion."

The current version of IHDPS is based on the 15 attributes listed in Figure 1. This list may need to be expanded to provide a more comprehensive diagnosis system. Another limitation is that it only uses categorical data. For some diagnosis, the use of continuous data may be necessary. Another limitation is that it only uses three data mining techniques. Additional data mining techniques can be incorporated to provide better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would definitely give better results. It is also necessary to test the system extensively with input from doctors, especially cardiologists, before it can be deployed in hospitals. [Access to the system is currently restricted to stakeholders.]

User Interface Of IHDPS:



Conclusion: A prototype heart disease prediction system is developed using three data mining classification modeling techniques. The system extracts hidden knowledge from a historical heart disease database. DMX query language and functions are used to build and access the models. Five mining goals are defined based on business intelligence and data exploration. The goals are evaluated against the trained models. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

REFERENCES:

[1] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[2] Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.

[3] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[4] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", Proc. of the
9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[5] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

[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] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[9] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

[10] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

[11] Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/en-us/virtuallabs/aa740409.aspx , 2007.

[12] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[13] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005

[14] Tang, Z. H., MacLennan, J.: "Data Mining with SQL Server 2005", Indianapolis: Wiley, 2005.

[15] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[16] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.

[17] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management.6(4),50, , 2002.

Date of submission: 18 October 2012

Date of provisional acceptance: 11 November 2012

Date of Final acceptance: 28 February 2013

Date of Publication: 03 April 2013

Source of support: Nil; Conflict of Interest: Nil

Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques such as Decision Trees, NaÃ¯ve Bayes, and Neural Network is implemented in [10] using .NET. heart disease and patient without heart disease. 1 (one) indicates Patient with heart disease and 0 (zero) indicates Patient without heart disease).Â Clustering Clustering is data mining techniques used to analyze data objects without referring back to a known label. It is the most useful and important techniques used for the process of discovery of data distribution. Clustering has two major types available. Research Article. Open Access. Heart Disease Diagnosis Using Data Mining Techniques. Ramin Assari1*, Parham Azimi2 and Mohammad Reza Taghva1. 1Department of IT Management, Allameh Tabatabaâ€™i University, Tehran, Iran 2Faculty of Mechanical Engineering and Industrial Engineering, Islamic Azad University, Qazvin, Iran.Â Every day, modern computer-based systems collect large amounts of data using automatic data record systems in the healthcare field where data mining can extract a valuable knowledge from them. The next section briefly explains heart disease and the application of data mining techniques in treating such diseases. Heart disease As the leading cause of death in the world, heart disease, according. Data mining is a technique that is performed on large databases for extracting hidden patterns by using combinational strategy from statistical analysis, machine learning and database technology. Further, the medical data mining is an extremely important research field due to its importance in the development of various applications in flourishing healthcare domain.Â [10] have carried out a research work and have built a model known as Intelligent Heart Disease Prediction System (IHDPS) by using several data mining techniques such as Decision Trees, NaÃ¯ve Bayes and Neural Network. Shantakumar, et al. [11] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back-Propagation.