PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE

N. Aditya Sundar[1], P. Pushpa Latha[2], M. Rama Chandra[3]

[1]*Asst.professor, CSE Department, GMR Institute of Technology, A.P, India,* ***aditya.sundar@gmail.com***
[2]*Asst.professor, CSE Department, GMR Institute of Technology, A.P, India,* ***pushpalatha.p@gmrit.org***
[3]*Asst.professor, CSE Department, GMR Institute of Technology, A.P, India,* ***ramachandra.m@gmrit.org***

Abstract

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. This paper describes about a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier).This system can answer complex "what if" queries which traditional decision support systems cannot. Using medical profile0073 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. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. Presently we are analyzing the performances of the two classification data mining techniques by using various performance measures.

Index Terms: *Naive Bayes, WAC, Classification techniques, CRISP, etc...*

----------***----------

1. INTRODUCTION

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems.

Most hospitals today employ 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?" This is the main motivation for this paper.

1.1. Data mining

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad defines data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database". Giudici defines it as "a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database". 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).

Each data mining technique serves a different purpose depending on the modeling objective. The two mostcommon modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

Naive Bayes or 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. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

In Weighted Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting capability. Weighted Associative Classifier (WAC) is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation.

1.2. Classification Concepts

Classification is a classic data mining task, with roots in machine learning. A typical application is: "Given past records of customers who switched to another supplier, predict which current customers are likely to do the same." This specific application is known as Churn Prediction, but there are very many other applications such as predicting response to a direct marketing campaign, separating good products from faulty ones etc.

The "Classification Problem" involves data which is divided into two or more groups, or classes. In our example above, the two classes are "switched supplier" and "didn't switch". The data mining software is asked to tell us which of the groups a new example falls into. So, we might train the software using customer records from the last year, divided into our two groups. We then ask the software to predict which of our customers we're likely to lose. Of course, to ensure we can trust the predictions, there is generally a testing or validation stage as well.

1.3. CRISP-DM Methodology

This system uses the CRISP-DM (cross industry standard process for data mining) methodology to build the mining models. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding phase focuses on understanding the objectives and requirements from a business perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives.
Data understanding phase uses the raw the data and proceeds to understand the data, identify its quality, gain preliminary insights, and detect interesting subsets to form hypotheses for hidden information.

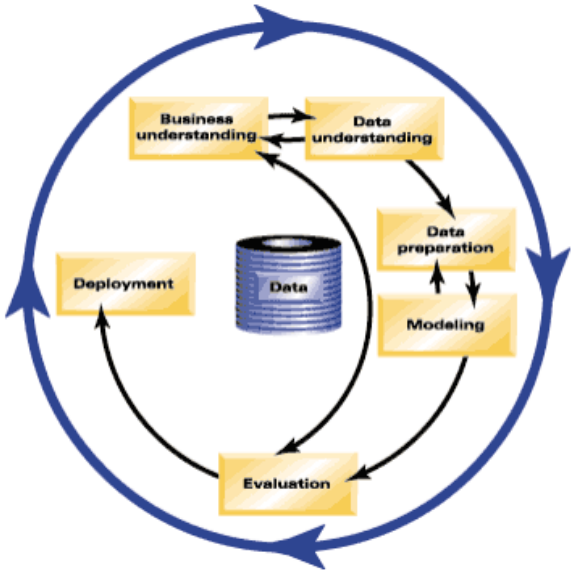

Fig-1: CRISP-DM Phases

Data preparation phase constructs the final dataset that will be fed into the modeling tools. This includes table, record, and attribute selection as well as data cleaning and transformation. The modeling phase selects and applies various techniques, and calibrates their parameters to optimal values. The evaluation phase evaluates the model to ensure that it achieves the business objectives. The deployment phase specifies the tasks that are needed to use the models.

Data Mining Extension (DMX), a SQL-style query language for data mining, is used for building and accessing the models' contents. Tabular and graphical visualizations are incorporated to enhance analysis and interpretation of results.

2. PROBLEM STATEMENT

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Identify the important Preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease."

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge- rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

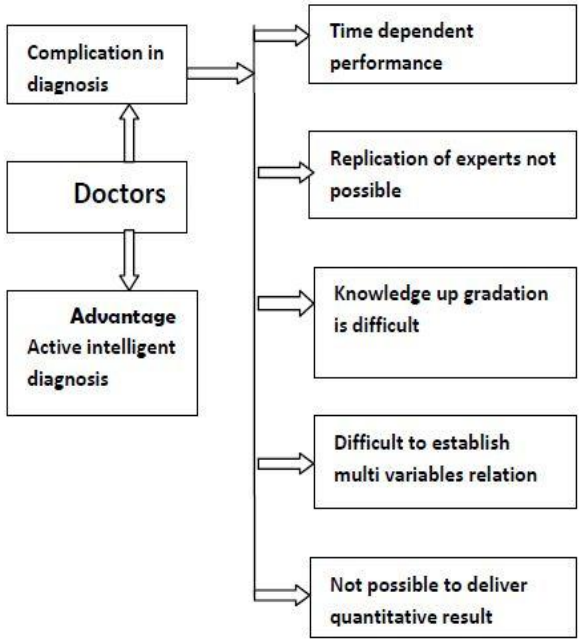

Fig-2: Traditional ways of making decisions

3. ALGORITHMS

3.1. Naive Bayes:

1. Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

P(Ci/X)= (P(X/Ci)P(Ci))/P(X)

3. As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

3.1.1. Pseudo code:

Calculate diagnosis="yes", diagnosis="no" probabilities Pyes, Pno from training input.

For Each Test Input Record

For Each Attribute
Calculate Category of Attribute Based On Categorical Division

Calculate Probabilities Of Diagnosis="Yes", Diagnosis="No" Corresponds To That Category P(Attr,Yes), P(Attr,No) From Training Input .

For Each Attribute

Calculate The Resultyes= Resultyes* P(Attr,Yes),Resultno= Resultno*P(Attr,No);

Calculate Resultyes= Resultyes *Pyes

Resultno= Resultno*Pno;

If(Resultyes > Resultno) Then Diagnosis="Yes";
Else Then Diagnosis ="No";

3.1.2. Formulae:

- Pyes=total number of yes/total number of records;
- Pno=total number of no/total number of records;
- P(attr,yes)=total number of yes in corresponding category/total number of yes;
- P(attr,no)=total number of yes in corresponding category/total number of yes;

3.2. Weighted association classifier:

It is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation. The major steps are as follows.

1) Initially, the heart disease data warehouse is pre -processed in order to make it suitable for the mining process.

2) Each attribute is assigned a **weight** ranging from o to 1 to reflect their importance in prediction model .Attributes that have more impact will be assigned a high weight(nearly 0.9)and attributes having less impact are assigned low weight(nearly 0.1)

3) Once the preprocessing gets over, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern. This algorithm uses the concept of Weighted Support and Confidence framework instead of tradition support and confidence. Rules generated in this step are known as CAR (Classification Association Rule) and is represented as X □ □ Class label where X is set of symptoms for the disease. Example of such rules are (Hypertension, "yes") _□ Heart_Disease="yes" and {(Age," >62"), (Smoking_habits,"yes"),(Hypertension,"yes")} □Heart_Disease="yes".

4) These rules will be stored in **Rule Base.**

5) Whenever a new patient's record is provide, the CAR rule from the rule base is used to predict the class label.

Weighted associative classifiers consist of training dataset T= $\{r_1, r_2, r_3 \ldots r_i \ldots\}$ with set of weight associated with each {attribute, attribute value} pair. Each i^{th} record r_i is a set of attribute value and a weight w_i attached to each attribute of ri tuple / record. In a weighted framework each record is set of triple {ai, vi, wi} where attribute ai is having value vi and weight wi, 0<wj<=1. Weight is used to show the importance of the item.

3.2.1. Attribute Weight*:*

Attribute weight is assigned depending upon the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In web mining visitor page dwelling time can be used to assign weight in medical domain symptoms can be assigned weight by expert doctor.

3.2.2. Attribute set weight:

Weight of attribute set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = (\sum_{i=1 \text{to } |X|} \text{weight } (a_i))/ \text{ Number of attributes in X}$$

3.2.3. Record weight/Tuple Weight:

Consider the data in relational table, the tuple weight or record weight can be defined as type of attribute weight. It is average weight of attributes in the tuple. If the relational table is having n number of attribute then Record weight is denoted by $W(r_k)$ and given by

$$W(r_k) = (\sum_{i=1 \text{to } |r_k|} \text{weight } (r_k))/ \text{ No. of attributes in a record}$$

3.2.4. Weighted Support:

In associative classification rule mining, the association rules are not of the form X–>Y rather they are subset of these rules where Y is the class label. Weighted support WSP of rule X–>Class_label, where X is set of non empty subsets of attribute-value set, is fraction of weight of the record that contain above attribute-value set relative to the weight of all transactions. This can be given as

$$WSP(X \text{–>} Class\_label) = (\sum_{i=1 \text{to } |r_k|} \text{Weight } (r_k))/ (\sum_{k=1 \text{ to } |n|} \text{weight } (r_k))$$

Here n is the total number of records.

3.2.5. Weighted Confidence:

Weighted Confidence of a rule X □–>Y where Y represents the Class label can be defined as the ratio of Weighted Support of (X –>Y) and the Weighted Support of (X).

Weighted Confidence =

Weighted support(x * y)/Weighted support(x)

3.2.6. Formulae:

Recordweight= summation of weights of the items /total number of items present;

Total weight= summation of the Recordweights of all records

W_support(X) =summation of the Record weights of all records which contains X /Totalweight;

W_confidence(X->Y)=W_support(XUY)/W_support(X);

3.2.7. Pseudo code:

Calculate Frequent itemsets from the items by using apriori algorithm;
For each frequent itemset
{
Calculate W_support(frequent itemset);
Claculate W_confidence(frequent itemset ->diagnosis_yes);

If(w_confidence >min_confidence)
Store it in the rulebase_yes with their
w_supports,w_confidences;
Claculate W_confidence(frequent itemset ->diagnosis_no);

If(w_confidence >min_confidence)
Store it in the rulebase_no with their
w_supports,w_confidences;
}

3.3. Apriori Algorithm:

3.3.1. Apriori property:

A subset of a frequent itemset must also be a frequent itemset i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset.
1. Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset).
2. Use the frequent itemsets to generate association rules.

Join Step: Ck is generated by joining Lk-1with itself
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

3.3.2. Pseudo-code:

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};

for (k = 1; Lk !=null; k++) do begin

C_{k+1} = candidates generated from L_k;

C_{k+1} = candidates generated from L_k;

C_{k+1} = candidates generated from L_k;

C_{k+1} = candidates generated from L_k;

C_{k+1} = candidates generated from L_k;

For each transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
end
return U L_k;

3.3.3. Data source

A total of 2268 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Figure 2 lists the attributes. The records were split equally into two datasets: training dataset (1857 records) and testing dataset (411 records).

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.

3.3.4. 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, value2: 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; value2: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 (value1: unsloping; value 2: flat; value 3: downsloping)

7. CA – number of major vessels colored by fluoroscopy (value 0 – 3)

8.Thal (value 3: normal; value 6: fixed defect; value7: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. PERFORMANCE EVALUATION

The effectiveness of models was tested using two methods: 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. Classification Matrix:

Classification Matrix displays the 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.

Count for Naive Bayes On Diagnosis Group		
Predicted	0(actual)	1(actual)
0	25	56
1	34	296

Count for weighted association classifier On Diagnosis Group		
Predicted	0(actual)	1(actual)
0	6	12
1	53	340

Fig-3. Results of Classification Matrix for all the two models

4.2. Lift Charts:

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 5 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 red, green and blue lines show the ideal, Naïve Bayes and WAC models respectively. The chart shows the performance of the models across all possible states. The model ideal line (red) is at 45-degree angle, showing that if 50% of the test dataset is processed, 50% of test dataset is predicted correctly. The chart shows that WAC gives (84%) followed by Naïve bayes (78%).

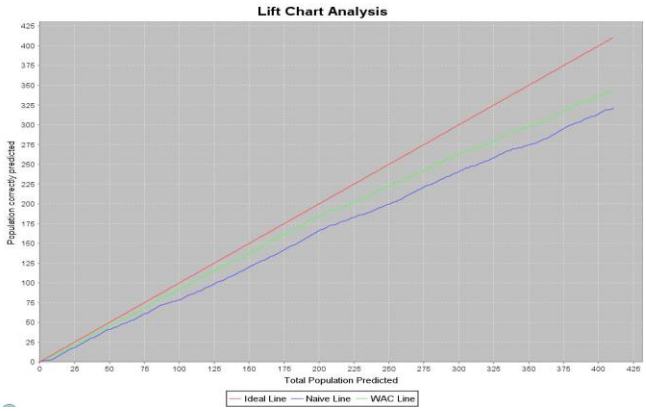

Fig -4: Lift charts analysis

4.3. Bar Charts:

Bar charts as shown in the figure5 actually how many records are taken for testing and out of those how many are with diagnosis "yes " and how many are with diagnosis "no" and after testing the result analysis in the same manner as shown in below figure5 . From the bar charts below we can say that out of 411 testing records for naive bayes 321 predicted correctly and 90 records predicted wrongly and for WAC 346 predicted correctly and 65 predicted wrongly.

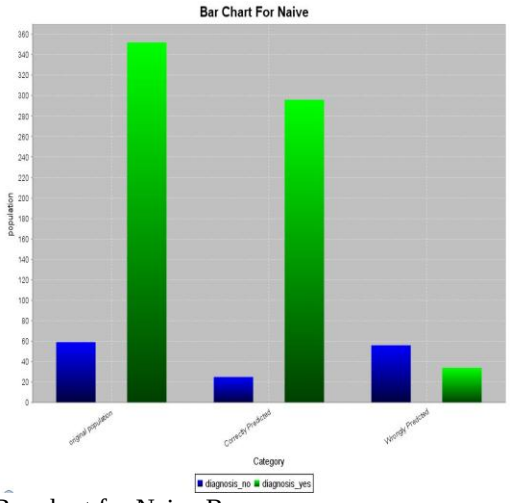

Fig-5: Bar chart for Naive Bayes

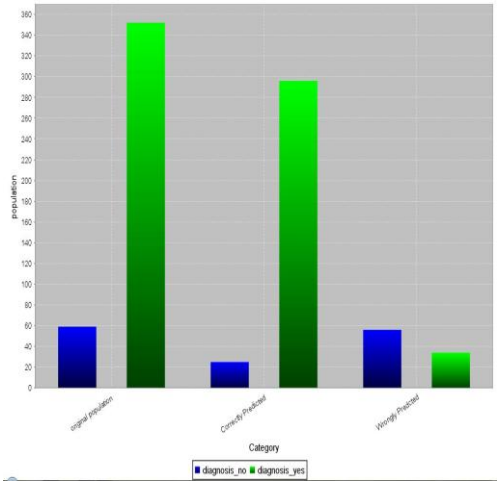

Fig-6: Bar chart for WAC

4.4. Pie charts:

The pie chart is perhaps the most widely used statistical chart in the business world and the mass media. Pie charts presented here can explain clearly what the performance level of each technique is.fig 7 shows pie charts for both Naïve Bayes and WAC techniques.

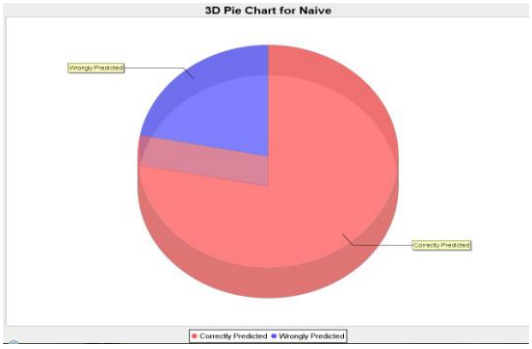

Fig-7 Pie chart for Naive Bayes

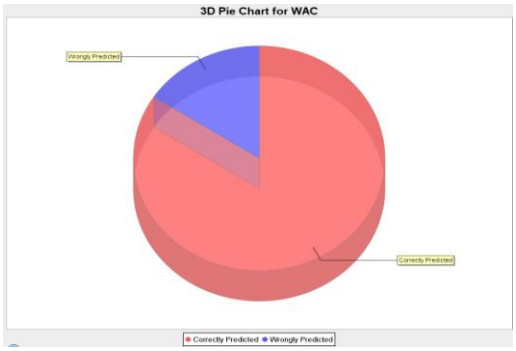

Fig -8 Pie chart for WAC

4.5. Output Screens

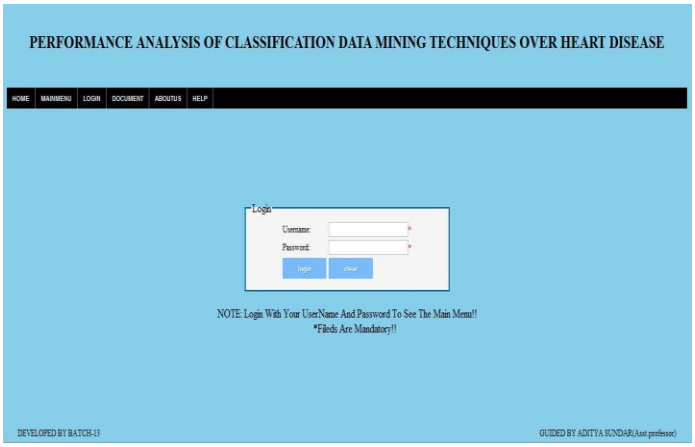

Fig-9: Login page

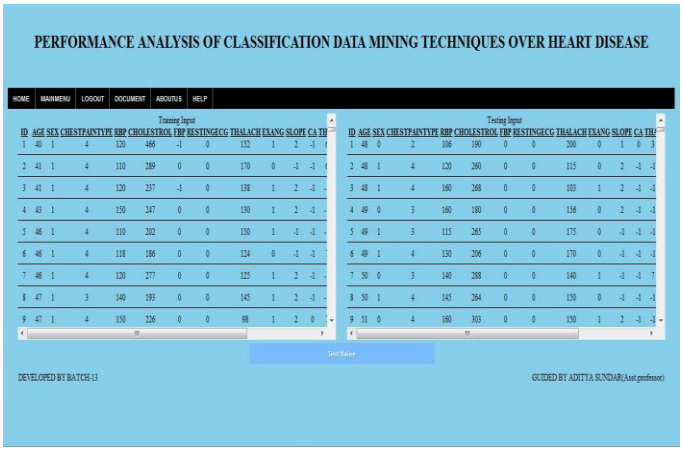

Fig-10: Naive home page

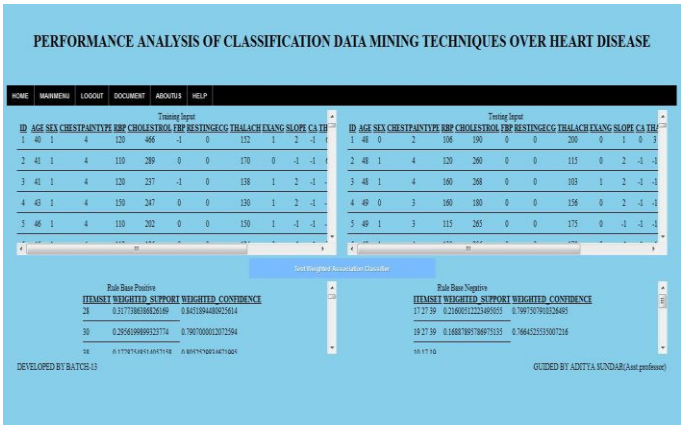

Fig-11: WAC home page

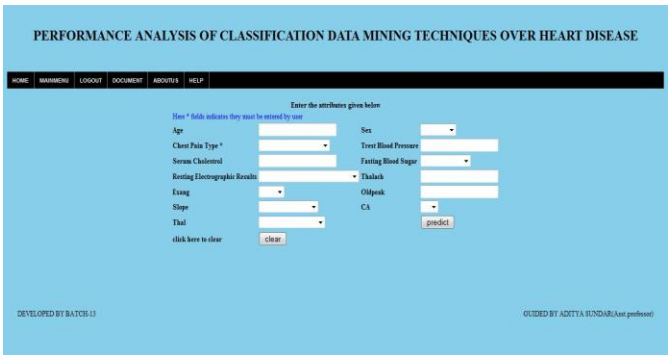

Fig-12: Data Inputs page

5. CONCLUSION

This system is developed using two 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. The models are trained and validated against a test dataset. Classification Matrix methods are used to evaluate the effectiveness of the models. The two models are able to extract patterns in response to the predictable state.

This system can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 15 listed in Figure 1. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data.

Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining.

Data Mining Extension (DMX) query language was used for model creation, model training, model prediction and model content access. All parameters were set to the default setting except for parameters "Minimum Support = 1" for Decision Tree and "Minimum Dependency Probability = 0.005" for Naïve Bayes. The trained models were evaluated against the test datasets for accuracy and effectiveness before they were deployed in HDPS. The models were validated using Classification Matrix.

5.1. Benefits and limitations

This system can serve a training tool to train nurses and medical students to d i a g n o s e patients with h e a r t 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 this system is based on the 15 attributes listed in Figure 3. 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 two 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.

REFERENCES

[1]. 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.

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

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

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

[5]. Han, J., Kamber, M.: *"Data Mining Concepts and Techniques"*, Morgan Kaufmann Publishers, 2006.

[6]. Ho, T. J.: *"Data Mining and Data warehousing"*, Prentice Hall, 2005.

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

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

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

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

[11]. Obenshain, M.K: *"Application of Data Mining Techniques to Healthcare Data"*, Infection Control and Hospital Epidemiology, 25(8), 690–695, 200.

BIOGRAPHIES:

N. Aditya Sundar is M.Tech in Computer Science from Andhra University, A.P., India. Since 2008 he has been working as assistant professor in the department of CSE.He is presently working as Assistant Professor in Department of Computer Science and Engineering, GMR Institute of Technology, Rajam, A.P, and India**.** His area of research includes Cloud Computing, Network Security, Data Mining, and Web Technologies. He can be reached at: **aditya.sundar@gmail.com.**

P.Pushpa Latha is M.Tech in Computer Science from Andhra University, A.P., India, presently pursuing her Ph.D in the same university. Since 2005 she has been working as assistant professor in the department of CSE. She is presently working as Assistant Professor in Department of Computer Science and Engineering, GMR Institute of Technology, Rajam, A.P, India**.** Her area of research includes Network Security, Data Mining, and Fuzzy logic. She can be reached at: **pushpalatha.p@gmrit.org.**

M. Ramachandra is M.Tech in Computer Science from NIT, Tiruchi, India. Since 2007 he has been working as assistant professor in the department of CSE, GMR Institute of Technology, Rajam, A.P, and India**.** His area of research includes Computer Networks, Network Security and Data Mining. He can be reached at: **rama00565@gmail.com**