Development of a Prototype for Critical Disease Predictions using Data Mining

Mohammad Taha Khan
Research Scholar, Suresh Gyan Vihar University
Mahal Jagatpura ,Jaipur,Rajasthan
e-mail: ertaha82@gmail.com

Professor Dr. Shamimul Qamar
Department of Computer Networks and Communication Engg.
College of Computer Science
King Khalid University,Abha, Ksa
e-mail: drsqamar@rediffmail.com.com

Dr. Ripu Ranjan Sinha
Associate Dean Research,
Suresh Gyan Vihar University,
Jaipur, India
e-mail: ripuranjan.sinha@mygyanvihar.com

Abstract—The goal of this paper is to present breast cancer prototype model along with the prediction of heart diseases by employing data mining techniques. The data used in the study had been retrieved from Public-Use Data, which is available online. The data comprised of 699 and 909 records for breast cancer and heart disease respectively. For data prediction and mining, C4.5 and C5.0, which are decision tree algorithms, were used on the data, used in the study. The results of both data sets using both algorithms were also compared. The paper also outlines the significance of evidence based medicine, which is the novel and innovative approach in healthcare decision making process [5]. It is essential that the clinical decisions are supported and based on scientific evidence, which ensures that they are sound and effective decisions. This paper also will depict the importance of data mining in modern healthcare.

Keywords*- Health care Prediction, data mining, EBM*

*****

I. INTRODUCTION

In healthcare and medical sector, there is need to provide accurate and precise diagnosis and treatment to patients in order to meet their requirements and provide them high quality and affordable care. Literature suggests that the quality of the service provided to the patients in healthcare organization refers to accurate diagnosis and providing them treatments that are instrumental in treating it efficiently [8]. Furthermore, hospitals and healthcare organizations also focus on reducing their costs and therefore, they aim at reducing costs associated with clinical testing by using computer based systems and decision support systems. Majority of the hospitals have adopted hospital information systems to store, record and manage patient data [7]. These systems have large amount of data, which is stored in the form of images, text, charts and numbers. Furthermore, such data has hidden information, which can be extracted from the huge whirlpool of information to help clinical decision making. The goal of this research is to investigate how data can be extracted into useful information and can help healthcare practitioners in their decision making.

The use of efficient and effective prediction system designed specifically for cancer and heart disease can be instrumental in improving the effectiveness of the healthcare organizations and can aid clinicians to take strong and effective decisions. It is essential that clinicians and patients are aware of the dangers of fatal diseases such as cancer and heart diseases and therefore, require efficient treatment. Consequently, modern healthcare organizations have adopted data mining techniques. Data mining techniques are used effectively to improve the efficiency of classification and prediction systems and therefore, can aid medical practitioners in their decision making process. This can be beneficial in improving the quality of care for patients, while it can reduce operational costs and thus, can lay out the foundation for further clinical studies.

The goal of this research is to use the dataset on heart disease and breast cancer, which is available to the public by using C4.5 and C5.0 classification algorithms for prediction, analysis of the data and compare the results of both.

The paper has been outlined as: section 2 discusses the significance of data mining in the healthcare sector, section 3 discusses the C4.5 and C5.0 classification algorithms, section 4 presents the data mining case studies of heart disease and cancer prediction, section 5 discusses the prototype model and section 6, is the final section, which provides conclusion and direction for future research.

II. IMPORTANCE OF DATA MINING IN HEALTHCARE

The use of information and communication technology in the healthcare industry had increased significantly.
Consequently, medical databases usage in healthcare organizations has increased to manage patient data and information in an efficient manner. The use of technology is the primary motivator for researchers and professionals to adopt information technology and decision support systems. The storage of data increases and consequently, data mining techniques can be used to efficient in extracting hidden information and extracting knowledge, which can be used to improve the quality of care provided to patients, while reducing the costs. The use of data mining technique can be used to address several questions [6]:

- How patient treatment can be improved for patients through analysis of patient data?
- How patient records can be used to make sound clinical decisions?
- How patient data and records can be used to treat cancer patients? Should the treatment comprise of chemotherapy or radiation therapy alone or together?

The goal of using data mining is to recognize the pattern and thus, it aims at discovering the patterns of the data, which cannot be detected using conventional statistical methods. Data mining techniques and procedures are based on several concepts such as statistical analysis, machine learning and visualizing. The best model that is based on data characteristics is applied by the data mining algorithms. Data mining models are classified into two categories [6], which are discussed as follows:

Descriptive Models are used to identify the data patterns with the help of rules of association, visualization, clustering techniques and pattern recognition.
Predictive Models are used for predicting the data. For instance, for diagnosing a disease, the predictive modelling can be used to determine the number of symptoms that are present in other patients along with the viable treatment options. Prediction uses classification and regression analysis of the data along with time series analysis. Research suggests that classification is the main and fundamental aspect of predictive modelling. Fig.1 shows the significance and importance of data mining in modern healthcare practice.

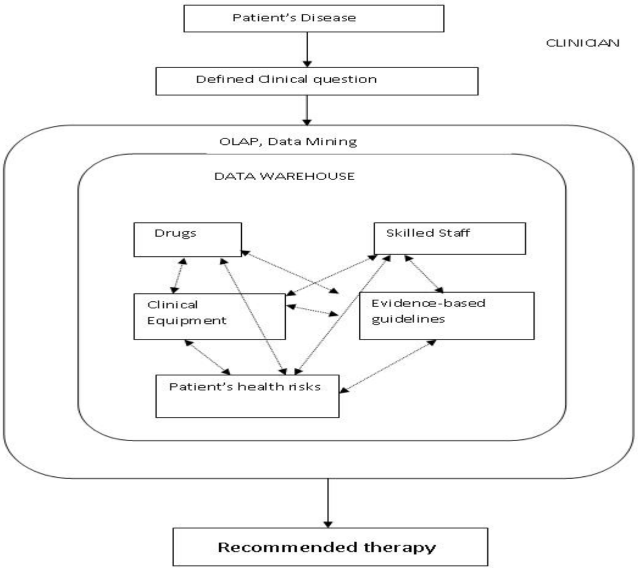


Fig.1 use of data mining in better health delivery [4]

III. CLASSIFICATION TECHNIQUES IN DATA MINING

Classification is considered to be an important aspect of predictive modelling and it aims at finding the model from the given set of data. It is considered to be a mapping strategy, which maps the data into a particular class, which has been defined already. This model has to follow a set of rules, which are dependent on the characteristics of the data that is under review. These rules are also applicable on the classification of data items that are unknown and would be detected in the future. It is considered to be one of the most important and essential technique related to data mining. The use of classification model in medical diagnosis can be used efficiently. For instance, diagnosing the symptoms of new patients can be achieved using the classification model with the already known cases that are available.

Classification needs to address the following:

Function; f = D→ C where each ti € D is mapped to f (ti) belonging to some Cj [3].
Where:
1. D is a database of patients with tuples (x1, x2 … xn)
2. x1, x2 … xn are values of attributes A1,A2 … .An relevant to a particular disease.
3. C= {C1, C2 … Cn} is set of classes of disease depending on its severity.

Classification model can be applied using the decision tree, which is considered to be an important aspect of discovering data and discovering the knowledge. The decision tree model is considered to be based on predictive modeling and thus, focuses on classifying the data in the form of tree. Complex and descriptive trees are called regression trees. They are also known as classification trees. For decision trees, training and testing data is required. The former is used in constructing data trees and thus, is part of the huge pool of data. The testing data is used for determining the accuracy of the data. It also analyzes whether the data has been misclassified in the decision tree. Consequently, it is associated with the reliability and validity of the data.

A. *Classification Algorithms C4.5 and C5.0*

C4.5 algorithm: [14] C4.5 is a classification algorithm, which has the ability to create decision trees by using training data. The formation of the decision trees are based on information entropy. Each of the characteristic of the data is selected by the C4.5, which is then divided into subsets, belonging to a particular class. The criteria of C4.5 is normalized information gain, which allows it to select the attribute and split the data accordingly. The attribute that has the highest normalized information gain is considered to be the criteria for making the decision. Consequently, after the decision is made, the small sublists are made. The C4.5 algorithm has some cases, which are discussed as follows:

1. The samples that are in the same class in the list, then, a leaf node is created for the decision tree. This is done to select the class. .
2. Information gain is not provided by the features. C4.5 is responsible for creating the decision node by using the class that has high expected value.
None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
3. Expected value is used by C4.5 to create the decision node higher if an unseen class has been encountered.

i). Tree Generation:

Entropy and Gain is used in creating the tree.

$$I(P) = -\sum_{i=1}^{k} p_i * \log(p_i)$$

Where p_i is the proportion of instances in the dataset that take the i_{th} value of the target attribute.

Gain is:

$$Info(X,T) = \sum_{i}^{n} \frac{|T_i|}{|T|} Info(T_i)$$

Where i is a value of X, |Ti| is the subset of instances of T where X takes the value i, and |T| is the number of instances

ii). Pruning Trees.

To reduce the error and to improve accuracy and to avoid overfitting, pruning algorithm is utilized. Pruning tree refers to the technique that is used in substituting the entire subtree by a leaf. If the subtree had expected error rate that is higher than the single lead, then the replacement is occurs. In this research, we would create the classification tree and use pruning for simplification.

B. C5.0 Algorithm

C5.0 and C4.5 algorithm have identical pseudo code. However, they are both different. The improvements made in C5.0 as compared to C4.5 are listed as follows:

1. Speed - C5.0 is quicker and faster than C4.5 in terms of magnitude.

2. Memory usage - C5.0 is more efficient in terms of memory

3. Smaller decision trees - C5.0 has the tendency to produce approximately the same results as C4.5 for small decision trees.

4. Support for boosting – C5.0 has improved boosting, which increase its accuracy and efficiency.

5. Weighting - C5.0 has the ability to weigh the data based on different attributes.

6. Winnowing - C5.0 has the ability to winnow the data as compared to C4.5

IV. Case Studies of Cancer and Heart Disease Prediction

This section of the research focuses on the case studies used: prediction of breast cancer and prediction of heart diseases.

A. Case Study 1: Breast Cancer Prediction

According to experts, breast cancer is the leading cause of death among in women all over the world [19]. In developing countries such as India, Pakistan and Bangladesh, the incidence of breast cancer among women is on the rise and is considered to be the primary cause of death among woman. According to the data compiled by Indian Council of Medical Research (ICMR), breast cancer is a serious problem in India. Reports suggest that it prevails in urban and rural dwellings. Research suggests that one out of twenty two women are most likely to suffer from breast cancer [12], while in America with one in eight being a victim of this deadly cancer.
University of Wisconsin Hospitals, Madison (Dr. William H. Walberg) [16] is having dataset for breast cancer online .This online available is used for the breast cancer prediction case study.

i). Breast Cancer data Attributes:

Total Cases:	**599**
Attribute	**Domain**
1. Sample code number	id number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	(2 for benign, 4 for malignant)

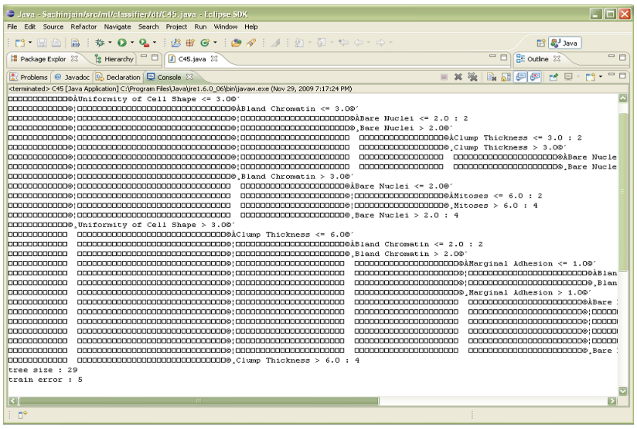
ii). Specification of Attributes:

The target attribute:
Class
Sample code number:
ignore
Clump Thickness:
continuous
Uniformity of Cell Size:
continuous
Uniformity of Cell Shape:
continuous
Marginal Adhesion:
continuous
Single Epithelial Cell Size:
continuous
Bare Nuclei:
continuous
Bland Chromatin:
continuous
Normal Nucleoli:
continuous
Mitoses:
continuous

The target attribute is class which can have two values either 2(Benign) or 4(Malignant).Malignant is cancerous.
Malignant tumors can invade and destroy nearby tissue and spread to other parts of the body Benign is not cancerous. Benign tumors may grow larger but do not spread to other parts of the body. Value to class attribute is given 2 and 4 to avoid the conflict with the values of other attributes. There are several attributes mentioned above which can have value from1 to 10.C 4.5 and C5.0 Programs supports three type of files: Names files Provides names for classes, attributes, and attribute values, Data file describe the training cases for generating the decision tree and/or and test file used to evaluate the produced classifier.

iii). Decision Tree and Rules Generated:

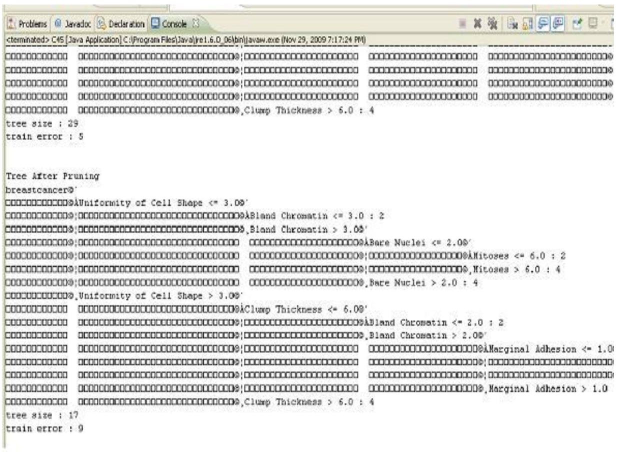
Following Fig.2 depicts the tree generated using c4.5 algorithm. Tree size is 29 with 5 train error.5 train errors means after running the 400 records on C4.5 there are five cases where error was noted down.


Java - Sachinjain/src/ml/classifier/dt/C45.java - Eclipse SDK
File Edit Source Refactor Navigate Search Project Run Window Help
Java
Package Explor
Hierarchy
C45.java
Outline
Problems
Javadoc
Declaration
Console

```
<terminated> C45 [Java Application] C:\Program Files\Java\jre1.6.0_06\bin\javaw.exe (Nov 29, 2009 7:17:24 PM)
□□□□□□□□□□□□⊚ÀUniformity of Cell Shape <= 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚ÀBland Chromatin <= 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚ÀBare Nuclei <= 2.0 : 2
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚¸Bare Nuclei > 2.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□⊚ÀClump Thickness <= 3.0 : 2
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□⊚¸Clump Thickness > 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚ÀBare Nucle
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¸Bare Nucle
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Bland Chromatin > 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚ÀBare Nuclei <= 2.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□⊚ÀMitoses <= 6.0 : 2
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□⊚¸Mitoses > 6.0 : 4
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¸Bare Nuclei > 2.0 : 4
□□□□□□□□□□□□⊚¸Uniformity of Cell Shape > 3.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚ÀClump Thickness <= 6.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚ÀBland Chromatin <= 2.0 : 2
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚¸Bland Chromatin > 2.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚ÀMarginal Adhesion <= 1.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□⊚ÀBlan
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Blan
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¸Marginal Adhesion > 1.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚ÀBare
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚¸Bare
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Clump Thickness > 6.0 : 4
tree size : 29
train error : 5
```

Fig.2.Tree Generated before pruning using c4.5

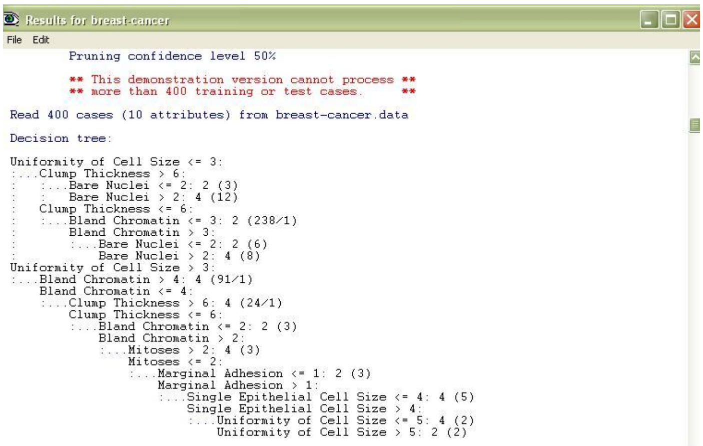
As pruning a tree is the action to replace a whole subtree by a leaf which reduces the size of tree. Following Fig.3 depicts tree after pruning. Tree size is 17.



```
<terminated> C45 [Java Application] C:\Program Files\Java\jre1.6.0_06\bin\javaw.exe (Nov 29, 2009 7:17:24 PM)
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□⊚
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Clump Thickness > 6.0 : 4
tree size : 29
train error : 5


Tree After Pruning
breastcancer⊚´
□□□□□□□□□□□□⊚ÀUniformity of Cell Shape <= 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚ÀBland Chromatin <= 3.0 : 2
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Bland Chromatin > 3.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚ÀBare Nuclei <= 2.0⊚´
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□⊚ÀMitoses <= 6.0 : 2
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□⊚¸Mitoses > 6.0 : 4
□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¸Bare Nuclei > 2.0 : 4
□□□□□□□□□□□□⊚¸Uniformity of Cell Shape > 3.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚ÀClump Thickness <= 6.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚ÀBland Chromatin <= 2.0 : 2
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□⊚¸Bland Chromatin > 2.0⊚´
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚ÀMarginal Adhesion <= 1.0
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¦□□□□□□□□□□□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□⊚¸Marginal Adhesion > 1.0
□□□□□□□□□□□□  □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□⊚¸Clump Thickness > 6.0 : 4
tree size : 17
train error : 9
```

Fig.3.Tree Generated after pruning using c4.5

Fig.4 shows the tree generated after running C5.0, which reads 400 cases with 10 attributes.



```
        Pruning confidence level 50%

        ** This demonstration version cannot process **
        ** more than 400 training or test cases.      **

Read 400 cases (10 attributes) from breast-cancer.data

Decision tree:

Uniformity of Cell Size <= 3:
:...Clump Thickness > 6:
:   :...Bare Nuclei <= 2: 2 (3)
:   :   Bare Nuclei > 2: 4 (12)
:   Clump Thickness <= 6:
:   :...Bland Chromatin <= 3: 2 (238/1)
:       Bland Chromatin > 3:
:       :...Bare Nuclei <= 2: 2 (6)
:           Bare Nuclei > 2: 4 (8)
Uniformity of Cell Size > 3:
:...Bland Chromatin > 4: 4 (91/1)
    Bland Chromatin <= 4:
    :...Clump Thickness > 6: 4 (24/1)
        Clump Thickness <= 6:
        :...Bland Chromatin <= 2: 2 (3)
            Bland Chromatin > 2:
            :...Mitoses > 2: 4 (3)
                Mitoses <= 2:
                :...Marginal Adhesion <= 1: 2 (3)
                    Marginal Adhesion > 1:
                    :...Single Epithelial Cell Size <= 4: 4 (5)
                        Single Epithelial Cell Size > 4:
                        :...Uniformity of Cell Size <= 5: 4 (2)
                            Uniformity of Cell Size > 5: 2 (2)
```

Fig.4.Rules Generated using c5.0

B. *Case Study 1I: Heart Disease Prediction*

Heart diseases are also one of the most deadliest diseases. Because of the life style now a days heart disease are becoming the very common. Prior knowledge of chances of getting a heart disease is very helpful for patient as well as clinicians for planning a better and effective treatment. This case is all about prediction of heart disease using the heart disease data set. The algorithms which are used again are C5.0 and C4.5. The purpose is to predict the presence or absence of heart disease given the results of various medical tests carried out on a patient.

We have used a total of 909 records with 75 medical attributes. This dataset is taken from Cleveland Heart Disease database [14].We have split this record into two categories: one is training dataset (455 records) and second is testing dataset (454 records). The records for each category are selected randomly. “Diagnosis” attribute is the target predictable attribute. Value “1” of this attribute for patients with heart disease and value “0” for patients with no heart disease. “PatientID” is 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.

i). Attribute Information:

----

1. Age (age in years)
2. Sex (1 = male; 0 = female)
3. Chest pain type (4 values)
    -- Value 1: typical angina
    -- Value 2: atypical angina
    -- Value 3: non-anginal pain
    -- Value 4: asymptomatic
4. Resting blood pressure
5. Serum cholesterol in mg/dl
6. Fasting blood sugar > 120 mg/dl   (1 = true; 0 = false)
7. Resting electrocardiography results (values 0, 1, 2)
    -- Value 0: normal
    -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Maximum heart rate achieved
9. Exercise induced angina (1 = yes; 0 = no)
10. Old peak = ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
    -- Value 1: upsloping
    -- Value 2: flat
    -- Value 3: downsloping
12. Number of major vessels (0-3) colored by flourosopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

ATTRIBUTES TYPES

Real: 1, 4,5,8,10,12
Ordered: 11,
Binary: 2, 6, 9
Nominal: 7,3,13
Variable to be predicted

-----------------------

Absence (1) or presence (2) of heart disease

ii). Decision Tree Rules Generated By C5.0

See5 [Release 2.06] Sat Nov 21 19:36:52 2013

Read 150 cases (13 attributes) from heartdisease.data

DECISION TREE:

```
Thal > 6:
:...ChestPain > 3: 2 (32/2)
:   ChestPain <= 3:
:   :...STSlope <= 1: 1 (8/2)
:       STSlope > 1: 2 (12/3)
Thal <= 6:
:...OldPeak > 2.8: 2 (6)
    OldPeak <= 2.8:
    :...ChestPain <= 3: 1 (60/6)
        ChestPain > 3:
        :...Vessels <= 0: 1 (23/6)
            Vessels > 0: 2 (9/1)
```

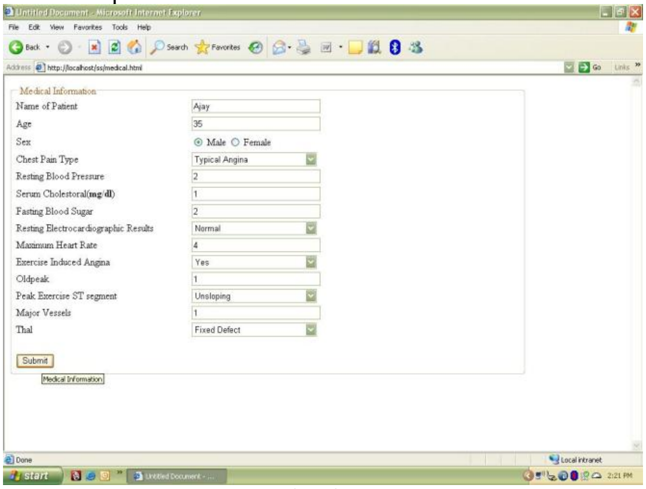
RULES:

```
Rule 1: (60/6, lift 1.6)
        ChestPain <= 3
        OldPeak <= 2.8
        Thal <= 6
        ->  class 1  [0.887]

Rule 2: (51/5, lift 1.6)
        ChestPain <= 3
        STSlope <= 1
        ->  class 1  [0.887]

Rule 3: (65/9, lift 1.5)
        OldPeak <= 2.8
        Vessels <= 0
        Thal <= 6
        ->  class 1  [0.851]

Rule 4: (27/1, lift 2.1)
        ChestPain > 3
        Vessels > 0
        ->  class 2  [0.931]

Rule 5: (32/2, lift 2.0)
        ChestPain > 3
        Thal > 6
        ->  class 2  [0.912]

Rule 6: (31/3, lift 2.0)
        STSlope > 1
        Thal > 6
        ->  class 2  [0.879]


Rule 7: (6, lift 2.0)
        OldPeak > 2.8
        Thal <= 6
        ->  class 2  [0.875]


Default class: 1
```

```
Evaluation on training data (150 cases):

        Rules
   ----------------
     No      Errors
      7   20(13.3%)   <<


     (a)   (b)    <-classified as
     ----  ----
      77     6    (a): class 1

      14    53    (b): class 2
```

C. *Working Prediction Model for Cancer*

As part of our project we have designed a working model for cancer/heart disease prediction. This model will predict the breast cancer's or heart disease class based on the rules created by C4.5 and C5.0 algorithms.Fig.6 shows the interface for input, which take Medical profiles of a patient such as age, sex, blood pressure and blood sugar etc as input and it can predict about presence or absence of cancer/heart disease.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://localhost/ss/medical.html
Done
start
Untitled Document - ...

Medical Information

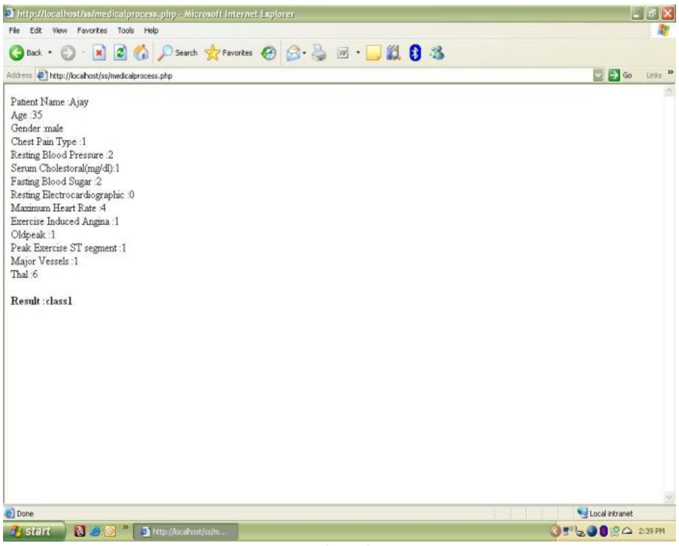
Name of Patient	Ajay
Age	35
Sex	● Male ○ Female
Chest Pain Type	Typical Angina
Resting Blood Pressure	2
Serum Cholestoral(mg/dl)	1
Fasting Blood Sugar	2
Resting Electrocardiographic Results	Normal
Maximum Heart Rate	4
Exercise Induced Angina	Yes
Oldpeak	1
Peak Exercise ST segment	Unsloping
Major Vessels	1
Thal	Fixed Defect

Submit

Medical Information

Fig.5.Interface for input

Fig7. Bellow shows a particular case belonging to class1 for heart disease.


http://localhost/ss/medicalprocess.php - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://localhost/ss/medicalprocess.php
Done
Local intranet
start
http://localhost/ss/m...

Patient Name :Ajay
Age :35
Gender :male
Chest Pain Type :1
Resting Blood Pressure :2
Serum Cholestoral(mg/dl):1
Fasting Blood Sugar :2
Resting Electrocardiographic :0
Maximum Heart Rate :4
Exercise Induced Angina :1
Oldpeak :1
Peak Exercise ST segment :1
Major Vessels :1
Thal :6

Result :class1

Fig.6.Interface for output

V. IMPORTANCE OF OF OPEN SOURCE SOFTWARE IN HEALTHCARE

Providing quality services at affordable prices is major challenge healthcare organizations (hospitals and medical centres).In term of ICT infrastructure hardware and software are capital goods for an organization.

Less price of a software means ICT is available at lower cost. This helps an organization to add to its resources and improves its process.

Open Source software is an important and growing class of software.

Open Source software is distinguished not by programming language, operating environment, nor application domain, but rather by the license(s) that governs the use, distribution, and, most importantly, the rights to access and modify the software's source code [21].

The philosophy of open source permits users to use, change, and improve the software, and to redistribute it in modified or unmodified forms. Together, software source code, licensing, and community have dramatically changed many conventional assumptions about software and the software industry itself. Acceptability of open source software is increasing day by day. Some of the reasons for using open source software include low total cost of Ownership, lack of software piracy issues, and availability of source code leading to high degree of customizability and scalability and extensive support freely available on Internet. When the source code of a program is available anyone can contribute by improving the code, adding new features, correcting errors, etc.

Healthcare is one of the important sectors for the economy of any developing country; if we get low cost ICT solutions for healthcare it is very beneficial for economic growth. Open source software have potential to be a key player for low cost quality healthcare delivery. Care2x, OpenVista, OpenEMR are some of free and open source healthcare software worldwide used.

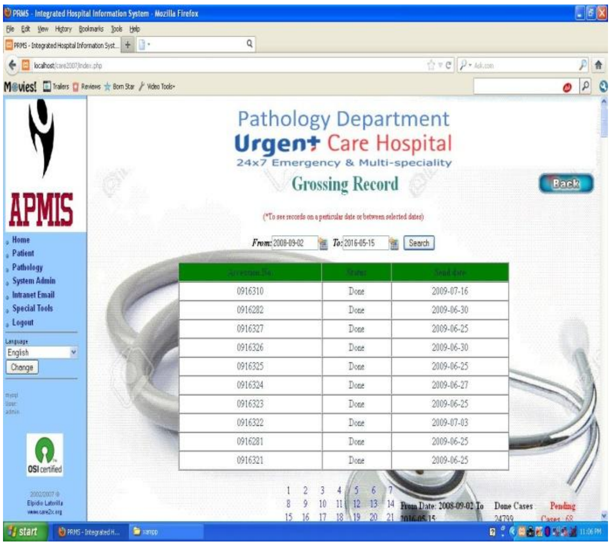
VI. MODIFIED CARE2X ADVANCED PATHOLOGY MANAGEMENT SYSTEM BASED (APMS)

We have developed one Advance Pathology Management System based on Care2x for Pathology of UrgentCare Hospital. UrgentCare is one of the premier hospitals in India with 160 beds. UrgentCare Pathology reports 100-150 cases per day. To improve the work process of this Pathology there is a requirement of an advance pathology management system. And maintaining the low cost was our primary goal. For this purpose we opted one existing open source software Care2x to customise it as per our requirement. CARE2X is an open source Web based Integrated Healthcare Environment (IHE) [22] under GNU/GPL. The project was started in May 2002 until today the development team has grown to over 100 members from over 20 countries. Its source code is freely distributed and available to the general public.

CARE2X [22] HIS is built upon other open-source projects: the Apache web server from the Apache Foundation the script language PHP [23] and the relational database management system mySQL [24]. CARE2X is modular and highly scalable so it is very easy to scale this application as per requirements. CARE2X is currently composed of four major components. Each of these components can also function individually. These components are HIS - Hospital/Health service Information System, PM - Practice (GP) management, CDS - Central Data Server, HXP - Health Xchange Protocol [22].

This advanced pathology management system is providing all features like Grossing, Sectioning, Reporting and Sample tracking with decision support.

Rules generated are used in this system to help the clinician in decision making.


PRMS - Integrated Hospital Information System - Mozilla Firefox
File Edit View History Bookmarks Tools Help
PRMS - Integrated Hospital Information Syst...
localhost/care2007/index.php
Movies! Trailers Reviews Born Star Video Tools
APMIS
Home
Patient
Pathology
System Admin
Intranet Email
Special Tools
Logout
Language
English
Change
OSI certified
2002/2007 © Elpidio Latorilla www.care2x.org
Back
start
PRMS - Integrated H...
xampp
11:06 PM

Pathology Department
Urgent Care Hospital
24x7 Emergency & Multi-speciality
Grossing Record

(*To see records on a particular date or between selected dates)

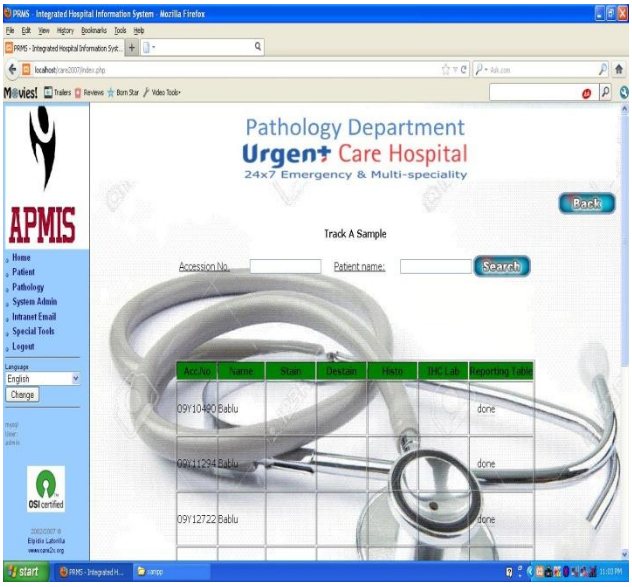
From: 2008-09-02 To: 2016-05-15 Search

Accession No	Status	Send date
0916310	Done	2009-07-16
0916282	Done	2009-06-30
0916327	Done	2009-06-25
0916326	Done	2009-06-30
0916325	Done	2009-06-25
0916324	Done	2009-06-27
0916323	Done	2009-06-25
0916322	Done	2009-07-03
0916281	Done	2009-06-25
0916321	Done	2009-06-25

1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21

From Date: 2008-09-02 To 2016-05-15: Done Cases 24799 Pending Cases: 68

Fig.7.Grossing option in APMS


PRMS - Integrated Hospital Information System - Mozilla Firefox
File Edit View History Bookmarks Tools Help
PRMS - Integrated Hospital Information Syst...
localhost/care2007/index.php
Movies! Trailers Reviews Born Star Video Tools
APMIS
Home
Patient
Pathology
System Admin
Intranet Email
Special Tools
Logout
Language
English
Change
OSI certified
2002/2007 © Elpidio Latorilla www.care2x.org
Back
start
PRMS - Integrated H...
xampp
11:03 PM

Pathology Department
Urgent Care Hospital
24x7 Emergency & Multi-speciality

Track A Sample

Accession No. Patient name: Search

Acc.No	Name	Stain	Destain	Histo	IHC Lab	Reporting Table
09Y10490	Bablu					done
09Y11294	Bablu					done
09Y12722	Bablu					done

Fig.8.Sample tracking in APMS

At the time of reporting system prompt the suggestions based on the rules. Based on the sample symptoms suggestion populated. This system is a step towards the evidence based medicine.

VII. CONCLUSIONS AND FUTURE WORK

If we talk about performance of these two algorithms, C5.0 handles missing values easily but C4.5 shows some errors due to missing values. Over running the dataset of breast cancer of 400 records C4.5 shows 5 train error whereas C5.0 show only

3 train errors. C5.0 produces rules in a very easy readable form but C4.5 generates the rule set in the form of a decision tree.

Data mining techniques play an important role in finding patterns and extracting knowledge from large volume of data. It is very helpful to provide better patient care and effective diagnostic capabilities. Evidence Based Medicine (EBM) is a new direction in modern healthcare.

EBM is as an important approach to make clinical decisions about the care of individual patients. This decision about patient is based on the best available Evidence. Its task is to prevent, diagnose and medicate diseases using medical evidence. It is all about providing best evidence, at right time in right manner to the clinician. External evidence-based knowledge cannot be applied directly to the patient without adjusting it to the patient's health condition. If the rules generated by this system is approved by medical experts that can be used as evidence for further use.

CARE2X is flexible generic multi-language open-source project. CARE2X is a very feature rich HIS, fully configurable for any clinical structure. After customization, it has the potential to become functional software to support workflows of Indian hospital. Efforts were made to explore the possibility of providing a low cost solution to Indian hospitals.

References

[1] Jaree Thongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang 'Breast Cancer Survivability via AdaBoost Algorithms'*HDKM,2008,wollongon,australia.*

[2] Diana Dumitru 'Prediction of recurrent events in breast cancer using the Naive Bayesian classification' Annals of University of Craiova, *Math. Comp. Sci. Ser.Volume 36(2), 2009, Pages 92-96 ISSN: 1223-6934.*

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

[4] Nevena Stolba and A Min Tjoa "The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making" December 24, 2005.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", J Healthcare Information Managment. 16(4), 50-55, 2002.

[6] Siri Krishan Wasan, Vasudha Bhatnagar and Harleen Kaur*The impact of data mining techniques on medical diagnostics" *Data Science Journal, Volume 5, 19 October 2006".*

[7] Herbert Diamond, Michael P. Johnson, Rema Padman, Kai Zheng, "Clinical Reminder System: A Relational Database Application for Evidence-Based Medicine Practice " INFORMSSpring National Conference, Salt Lake City, Utah- April 26, 2004.D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[8] Sellappan Palaniappan , Rafiah Awang "Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques" Proceedings of iiWAS2007.

[9] Infectious Disease Informatics and, outbreak detection,Daniel Zeng1, Hsinchun Chen, Cecil Lynch, Millicent Eidson, and Ivan Gotham.

[10] AMPATH Medical Record System AMRS): Collaborating toward An EMR for Developing Countries Burke W. Mamlin, M.D. and Paul G. Biondich, M.D., M.S.Regenstrief Institute, Inc. and Indiana University School of Medicine, Indianapolis, IN

[11] Global Epidemiological Outbreak Surveillance System Architecture:Ricardo Jorge Santos(1) and Jorge Bernardino

CISUC – Centre of Informatics and Systems of the University of Coimbra – University of Coimbra)ISEC – *Engineering Institute of Coimbra – Polytechnic Institute of Coimbra portugal*

[12] http://www.medindia.net/news/view_news_main.asp?x=7279

[13] Managing Diagnostic Process Data Using Semantic Web,Vili Podgorelec, Luka Pavlic Institute of Informatics, FERI, University of Maribor, Slovenia.Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07) 0-7695-2905-4/07

[14] http://en.wikipedia.org/wiki/C4.5_algorithm.

[15] ARIHITO ENDO, TAKEO SHIBATA, HIROSHI TANAKA 'Comparison of Seven Algorithms to Predict Breast Cancer Survival' *Biomedical Soft Computing and Human Sciences*, Vol.13, No.2, pp.11-16 (2008).

[16] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic) (Breast cancer dataset).

[17] http://archive.ics.uci.edu/ml/datasets/Heart+Disease (Heart Disease dataset).

[18] DMS Tutorial: http://dms.irb.hr/tutorial/tut_dtrees.php.

[19] ICMR Bulletin: http://www.icmr.nic.in/bufebruary03.pdf.

[20] Tipawan Silwattananusarn and Dr. KulthidaTuamsuk ' Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012' *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.*

[21] Michael Tiemann President Open Source Initiative Vice President Open Source Affairs, Red Hat November 1, 2009 'How Open Source Software Can Save the ICT Industry One Trillion Dollars per Year'.

[22] CARE2X; an Open Source Project. http://www.CARE2X.org .

[23] PHP An Open Source widely used language for web development, http://www.php.org .

[24] MySql Largest Open Source Database used by many renowne leading organizations http://www.mysql.com.