International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 8, October - 2012

Data Mining Classification Techniques Applied For Cancer Disease – A Case Study Using Xlminer

S. Jothi[1], S.Anita[2]
[1] Asst. Prof., Department of Computer Science, Jayaraj Annapackiam College for Women, Periyakulam, Theni Dt.
[2] Lecturer, Department of Electronics and Communication Engg., St. Anne's college of Engg. & Technology, Panruti, Cuddalore Dt.

ABSTRACT

Data mining techniques have been used in medical research for many years and have been known to be effective. In order to solve such problems as long-waiting time, congestion, and delayed patient care, faced by emergency departments, this study concentrates on building a hybrid methodology, combining data mining techniques such as association rules and classification trees. The methodology is applied to real-world emergency data collected from a hospital and is evaluated by comparing with other techniques. The methodology is expected to help physicians to make a faster and more accurate classification of cancer diseases.

Keywords **:** Data mining; Cancer Disease; XLMiner, Diagnosis; Prognosis; Classification.

I. INTRODUCTION

Cancer leads to approximately 25% of all mortalities, making it the leading cause of death in developed countries. Early and accurate detection of cancer is critical to the well being of patients. The most effective way to reduce cancer deaths is detect it earlier. This study paper summarizes various review and technical articles on cancer diagnosis and prognosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance cancer diagnosis and prognosis. This study paper entitled as *"*Data Mining Classification Techniques Applied for Cancer Disease*"* was done using *XLMiner*. The main objective of this paper is to find out which predict the various Classes for the people are affected by Cancer. For this data collected from several people. This is helps to classify, in which
category people are affected by which type of Cancer. Informations are collected based on Bone Cancer, Bladder Cancer, Stomach Cancer, Kidney Cancer and Uterus Cancer. It gives an overview of the current research being carried out on various cancer datasets using the data mining techniques to enhance the cancer diagnosis and prognosis. This process is done by using the data mining process and the classification algorithm.

II. PROBLEM DEFINITION AND DESCRIPTION

The data collected from individual person about their disease. We predicted various classes such as *Bone Cancer, Bladder Cancer, Stomach Cancer, Kidney Cancer and Uterus Cancer*. In my case study is to classify the disease belongs to the people's Symptoms. For example, Bone cancer have the symptoms of Pain in Bone, Nausea, Fracture in Bone, Vomiting. Kidney Cancer has the symptoms of Blood Pressure, Blood in Urine. Stomach Cancer has the symptoms of Nausea, Vomiting, Pain in abdomen, Blood in vomit. Uterus Cancer has the symptoms of Nausea, pain in abdomen, pain upon Urination. Bladder Cancer has the symptoms of blood in vomit, Blood in Urine, Pain upon Urination. In that my case study is to classify the disease belongs to the people's Symptoms.

III. KNOWLEDGE DISCOVERY AND DATA MINING

This section provides an introduction to knowledge discovery and data mining. We list the various analysis tasks that can be goals of a discovery process and lists methods and research areas that are promising in solving these analysis tasks.

3.1. The Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent

words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

1) ***Data cleaning:*** *A*lso known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.
2) ***Data integration*:** At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3) ***Data selection:*** At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4) ***Data transformation:*** Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5) ***Data mining:*** It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6) ***Pattern evaluation:*** This step, strictly interesting patterns representing knowledge are identified based on given measures.
7) ***Knowledge representation****:* Is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

3.2. *Data Mining Process*

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: *descriptive data mining tasks* that describe the general properties of the existing data, and *predictive data mining tasks* that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms.

Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects). Data mining involves some of the following key steps.

1) ***Problem definition:*** The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioural model.
2) ***Data exploration****:* If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
3) ***Data preparation:*** The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.
4) ***Modelling****:* Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
5) ***Evaluation and Deployment:*** Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

IV DATA MINING DEVELOPMENT PROCESS

4.1 ALGORITHM IMPLEMENTATION

In this paper, we have classified the disease according to the people's Symptoms using data mining techniques by implementing the algorithm. Classification is data mining technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "Sunny", "Rainy", or "Cloudy". Popular

classification techniques include Decision trees and neural networks. We need to use learning algorithms that can produce result in order to rank the testing examples. Algorithm which is used in classification called as decision tree. Decision tree methods build a collection of rules for use as a predictive model.

The advantage of this approach is that the rules are easy to understand, and they are frequently useful for discovering underlying business processes. The disadvantage of decision tree approach is that these models usually do not perform as well as other models. For then they developed a proprietary modification for standard decision tree algorithms.

Step 1:

Input variables are identified. They are declared as var1 as Sex, var2 as Age, var3 as Pain in Bone, var4 as Nausea, var5 as Fracture in Bone, var6 as Vomiting, var7 as Pain in Abdomen, var8 as Blood in Vomit, var9 as Blood Pressure, var10 as Blood in Urine,var11 as Pain Upon Urination and var12 as disease.

Step 2:

In this second step, to find out the probabilities for each attribute by using information gain. The information gain is calculated from the following formula,

$$\textbf{Info (D)} = - \{ \sum_{i=1}^{n} p_i \log_2 (p_i)$$

Where pi is the probability that an arbitrary tuple belongs to the class. Here all the classes are in the probability value of 1.5

Step 3:

Training log is used to find out the miss classify of the given class. It also used for growing the full tree using training date.

Step 4:

In this step, the full tree rules are used for find out the decision and terminal nodes. Where the decision nodes are 4, and the terminal nodes are 5.

Step 5: In this step, the tree was constructed with the corresponding attributes. The tree has taken the root node as the attribute Pain Upon Urination. The root node is selected by that high probability value. Finally, in this step to classify the actual class to predicted classes in given data such as Bone Cancer, Bladder Cancer, Stomach Cancer, Kidney Cancer and Uterus Cancer will be get from this step. The overall elapse time to run the XL Miner for this case study is 1Sec.

V. DATA MINING FINDINGS

The initial studies unveiled a number of relationships between variables as well as threshold values that justify further analysis. The several values of several attributes are useful predictors of retention and/or attrition. These explanations increase our confidence that the values of these attributes will continue to be predictors in the future. We have to classify how many persons are affected by Bladder Cancer, how many persons are affected by Bone Cancer, how many persons are affected by Kidney Cancer, how many persons are affected by Stomach Cancer, and how many persons are affected by Uterus Cancer by using XLMiner.

AGE	PAIN IN BONE	NAUSEA	FRACTURE IN BONE	VOMITING	PAIN IN ABDOMEN	BLOOD IN VOMIT	BLOOD PREASURE	BLOOD IN URINE	PAIN UPON URINATION	DISEASE
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER

SENIOR	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
MIDDLE AGED	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	YES	NO	NO	YES	YES	BLADDER CANCER
YOUNG	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	NO	KIDNEY CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	NO	NO	NO	YES	NO	YES	NO	NO	KIDNEY CANCER
MIDDLE AGED	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER

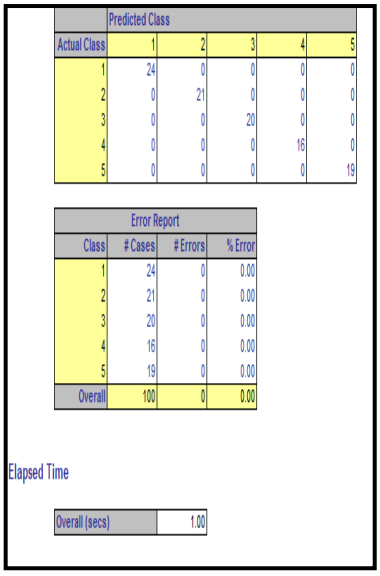
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	YES	NO	NO	YES	NO	NO	NO	YES	STOMACH CANCER

MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
MIDDLE AGED	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	NO	KIDNEY CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
MIDDLE AGED	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER

MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
SENIOR	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
YOUNG	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	YES	NO	YES	YES	YES	NO	NO	NO	STOMACH CANCER
MIDDLE AGED	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
SENIOR	NO	NO	NO	NO	NO	YES	NO	YES	YES	BLADDER CANCER
MIDDLE AGED	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
MIDDLE AGED	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER
SENIOR	NO	NO	NO	NO	NO	NO	YES	YES	NO	KIDNEY CANCER
YOUNG	NO	YES	NO	NO	YES	NO	NO	NO	YES	UTERUS CANCER
YOUNG	YES	YES	YES	YES	NO	NO	NO	NO	NO	BONE CANCER

Table 1 – Data Collection

Figure 1



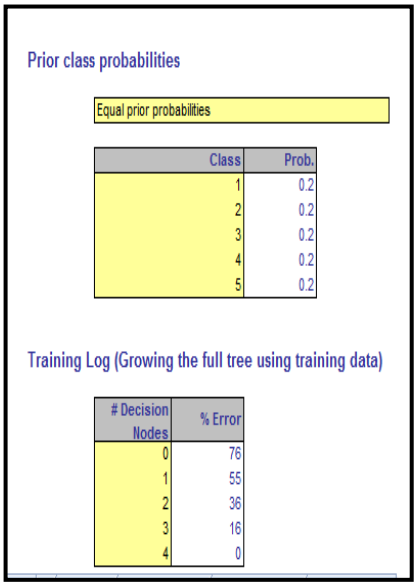
	Predicted Class				
Actual Class	1	2	3	4	5
1	24	0	0	0	0
2	0	21	0	0	0
3	0	0	20	0	0
4	0	0	0	16	0
5	0	0	0	0	19

Error Report			
Class	# Cases	# Errors	% Error
1	24	0	0.00
2	21	0	0.00
3	20	0	0.00
4	16	0	0.00
5	19	0	0.00
Overall	100	0	0.00

Elapsed Time

Overall (secs)	1.00

Figure 2



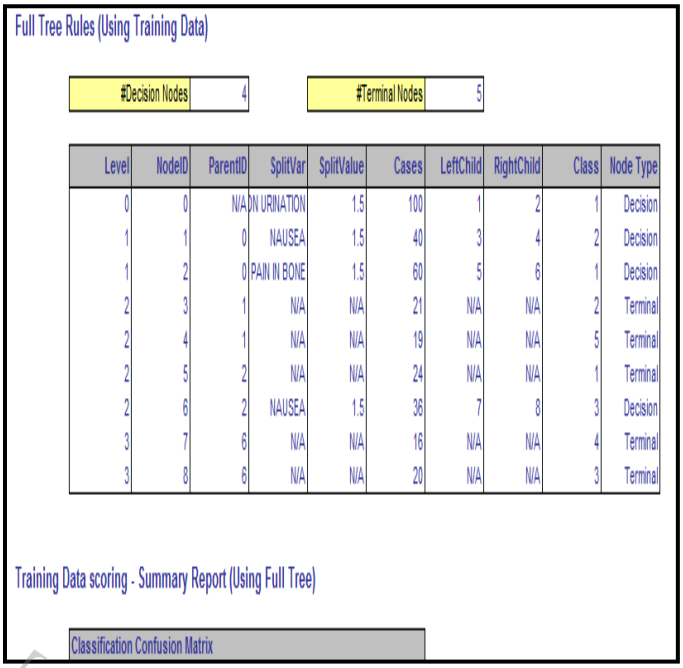
Prior class probabilities

Equal prior probabilities	
Class	Prob.
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

Training Log (Growing the full tree using training data)

# Decision Nodes	% Error
0	76
1	55
2	36
3	16
4	0

Figure 3: Prior Class Probabilities



Full Tree Rules (Using Training Data)

#Decision Nodes	4

#Terminal Nodes	5

Level	NodeID	ParentID	SplitVar	SplitValue	Cases	LeftChild	RightChild	Class	Node Type
0	0	N/A	ON URINATION	1.5	100	1	2	1	Decision
1	1	0	NAUSEA	1.5	40	3	4	2	Decision
1	2	0	PAIN IN BONE	1.5	60	5	6	1	Decision
2	3	1	N/A	N/A	21	N/A	N/A	2	Terminal
2	4	1	N/A	N/A	19	N/A	N/A	5	Terminal
2	5	2	N/A	N/A	24	N/A	N/A	1	Terminal
2	6	2	NAUSEA	1.5	36	7	8	3	Decision
3	7	6	N/A	N/A	16	N/A	N/A	4	Terminal
3	8	6	N/A	N/A	20	N/A	N/A	3	Terminal

Training Data scoring - Summary Report (Using Full Tree)

Classification Confusion Matrix

Figure 4 : Full Tree Rules

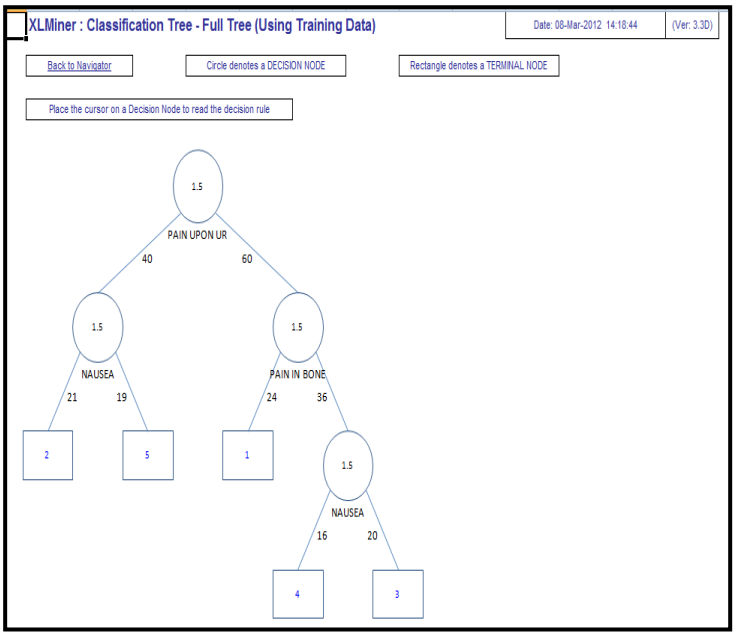


Figure 5 : Classification Tree – Full Tree

VI. CONCLUSION

The goal of classification is to build a set of models that can correctly predict the class of the different objects. The input to this method is set of objects (i.e., training data), the classes which these objects belongs to (i.e., dependent variables), and a set variables describing different characteristics of the objects (i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known a priori. Hereby, we collected all the data about which type of cancer are affect the people according to their Symptoms. The above technique can be successfully applied to the data sets for any cancer (such as Bone Cancer, Bladder Cancer, Stomach Cancer, Kidney Cancer, Uterus and so on), as it was successfully demonstrated on the bone, bladder, stomach, Kidney and ovarian cancer in this research. The early and accurate detection as well as classification of cancer using the XLMiner tool will help in the selection of treatment options.

BIBLIOGRAPHY

2. Agarwal R, Imielinski T, Swami AN. *"Mining Association rules between Sets of Items in Large Databases."* SIGMOD. June 1993, 22(2):207-16

3. Agarwal R, Srikant R,"*Fast Algorithms for mining Association Rules*", VLDB. Sep 12-15-1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

4. Mannila H,Toivonen H, Verkamo AI. "Efficiant Algorithms for Discovering Association Rules." AAAI workshop on Knowledge Discovery in Databases (SIGKDD). July 1994, Seattle, 181-92, ps.

5. Implementation of the algorithm in C# Retrieved from http://en.wikipedia.org/wiki/Apiriori_algorithm.

6. http://www.scribd.com/doc/28249613/Data-Mining-Tutorial

7. Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques," Software Technology and Engineering (ICSTE), 2nd International Conference, 2010, vol.2,pp.227-231.

8. Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011 189