International Journal of Computer Science and Engineering Open Access
Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

Disease Predication of Cardio- Vascular Diseases, Diabetes and Malignancy in Lungs Based on Data Mining Classification Techniques

Monali Dey[1*] and Siddharth Swarup Rautaray[2]

[1*,2]*Computer School of KIIT University, Bhubaneswar ,India*

www.ijcseonline.org

Received: 16/03/2014 Revised: 28/03/2014 Accepted: 22/04/2014 Published: 30/04/2014

Abstract— Data mining technology provides a user oriented approach to extract the hidden information from the large database. There are different algorithms used in data mining techniques like decision tree, Bayesian classifier, naive Bayes, neural network, , clustering etc. . Data mining in healthcare medicine deals with learning models to predict patients' disease. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The main goal of this paper is to analyze and implement the data mining algorithms using WEKA tool and comparison between c4.5 and Bayesian classifier.

Keywords— Bayesian Classification, Bayesian Networks, C4.5, Neural Network

I. INTRODUCTION

Computer Science is now getting more and more involved in the medicine and health sciences. The branch of computer science which is more actively and efficiently involved in medical sciences is Artificial Intelligence. Various healthcare Decision Support Systems have been constructed by the aid of Artificial intelligence. These systems are now widely used in hospitals and clinics. They are proved to be very useful for patient as well as for medical experts in making the decisions. Different methodologies are used for the development of those systems. The way of gathering the input data and to present output information's is different in different methodologies. Any computer program that helps experts in making healthcare decision comes under the domain of healthcare decision support system. An important characteristic of the Artificial Intelligence is that it can support the creation as well as utilization of the healthcare knowledge. Data mining is the core step, which results in the discovery of hidden and predictive information from large databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data"[1].Data mining involves six common classes of tasks: Anomaly Detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation. Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Classification – is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as "legitimate" or as "spam". Regression – Attempts to find a function which models the data with the least error. Summarization – providing a more compact representation of the data set, including visualization and report generation.

Corresponding Author: Monali Dey
Computer School of KIIT University, Bhubaneswar ,India

Data mining technology provides a user-oriented approach to extract hidden patterns from the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. . Advantages of Data mining: Predict future trends, customer purchase habits, Help with decision making, Improve company revenue and lower costs, Market basket analysis, Fraud detection. . Disadvantages: Great cost at implementation stage, Possible misuse of information, Possible inaccuracy of data.Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management [2].

1. Data modeling for health care applications
2. Executive Information System for health care

3. Forecasting treatment costs and demand of resources
4. Anticipating patient's future behaviour given their history
5. Public Health Informatics
6. e-governance structures in health care
7. Health Insurance

II.DATAMINING TECHNIQUES

Data mining technique is most important technique which is used in Knowledge Discovery in Database(KDD).KDD has different types of steps like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation etc. There are different types of techniques used in Data mining project. These include Decision tree, Bayesian networks, Naive Bayes, Neural networks etc.

A. Decision tree-It is the most frequently used techniques of data analysis. It is used to classify records to a proper class and is applicable in both regression and associations tasks.
In medical field decision trees specify the sequence of attributes. Such a tree is built of nodes which specify conditional attributes – symptoms X={x1,x2,.....xk},branches which show the values of S ih i.e. the *h*-th range for *i*-th symptom and leaves which present decisions Y={y1,y2.....yk} and their binary values Zdk={0,1} . A sample decision tree is presented in the fig1.[4][5]

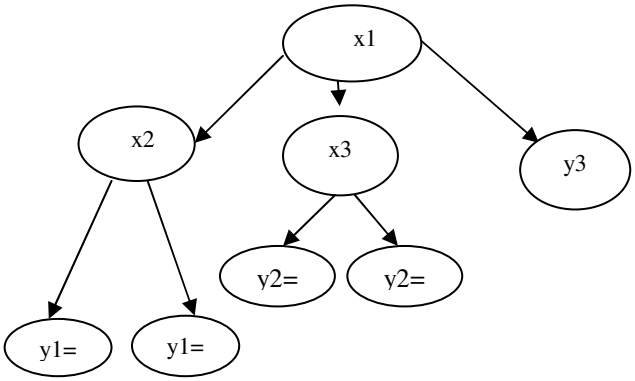

Fig 1 Decision tree applicable in medicine

B. Bayesian classifier- Bayesian classifier known as naive Bayesian classifier. Bayesian networks are graphical models, which allow the dependencies among subsets attributes. It is a simple probabilistic classifier, which is based on an assumption about mutual independency of attributes. The probabilities which is applied in the Naïve Bayes algorithm are calculated according to the Bayes Rule, the probability of hypothesis *H* can be calculated on the basis of the hypothesis *H* and evidence about the hypothesis *E* according to the following formula:[7]

$$P(H|E)= \frac{P(E|H)*P(H)}{P(E)} \qquad (1)$$

C. Neural Networks-In medical diagnosis the input to the neural network are the patient's symptoms the set X , and Y is the output of the diagnosis . There are 3 layers in neural networks: input layer, hidden layer, output layer. Hidden layer is the outcomes of the input layer..The condition between neurons have weights which is assigned to them.
Their values are calculated with the use of back propagation algorithm. In hidden layers there are some nonlinear features are added to the network..The out layer may have more than one output node which predict the different diseases. In a single neuron there are many input layers and one output layer. The input and output values are issued with the use of combination and activation function.

III.ANALYSIS OF CLASSIFICATION ALGORITHMS IN WEKA

Data mining algorithms which are indentified to be very common in HDSS are implemented in WEKA environment. The aim of this chapter is to familiarize one with the WEKA algorithms' implementation details, describe important parameters and show the ways of the result presentation. As WEKA is fully implemented in java programming languages, it is platform independent & portable. It is freely available under GNU General Public License. WEKA s/w contain very graphical user interface, so the system is very easy to access. There is very large collection of different data mining algorithms.[11][12]

A. C4.5
In WEKA environment the algorithm C4.5 is called J48 and it is the newest version of this algorithm's implementation. The parameters of C4.5 algorithm allows changing confidence threshold responsible for tree pruning, minimum number of instances which are permitted at a leaf. It is also possible to set the size of pruning set which is the number of data parts from which the last is used for tree pruning. Furthermore, WEKA's C4.5 decision tree may be pruned with the reduced error pruning. To achieve this it is essential to turn on reduced Error Pruning (set True instead default False). The generated decision tree may be presented in the text form. It is also possible to see graphical (more intuitive) form of the tree. The decision tree leafs have values in brackets like for instance (15.0/1.0) what means that 15 instances followed this formula correctly and 1 was misclassified.[6]*Advantages & disadvantages:* The advantages of the C4.5 are: Builds models that can be easily interpreted, Easy to implement, Can use both categorical and continuous values, Deals with noise. The disadvantages are: Small variation in data can lead to different decision trees (especially when the variables are close to each other in value) Does not work very well on a small training set.C4.5 is used in classification problems and it is the most used algorithm for building decision tree. It is suitable for real world problems as it deals with numeric attributes and missing values. The algorithm can be used for building smaller or larger, more accurate decision trees and the algorithm is quite time efficient.

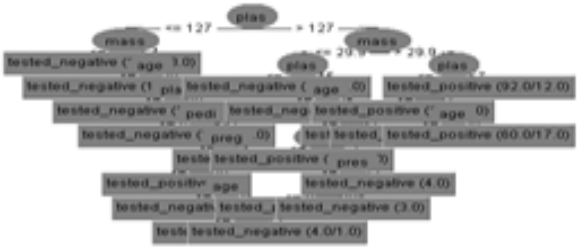

Figure 2 Graphical form of the C4.5 tree

Pseudocode of c4.5:

Input : Tree(t)

Step1: $f \rightarrow$ Class frequency[T]

step2: If [one class ∨ few classes]

Then R(leaf)

step3: Node(N)

Step3: ∀ Attribute(A)

$f \rightarrow$ Gain(A)

step4:N.test=Attribute(A)

step5:Find t_a

step6:$\forall T'(T)$

step7:ifT'=∅

child(N)=Leaf

else

step8:child(N)=T'

step9:f→ error (N)

return N

II. HYPOTHESIS

B. Bayesian classifier

Bayesian classifiers are statistical classifiers which predict class membership probabilities. comparing classification algorithms have found a simple Bayesian classifier known as naive Bayesian classifier. Bayesian networks are graphical models, which allow the dependencies among subsets attributes. It also be used for classification Naïve Bayes classifier has quite simple interface in WEKA environment . It allows one to select the kernel estimator for numeric attributes rather than a normal distribution and used Supervised Discretization while converting numeric attributes to normal ones. [7][8]

Pseudocode of Bayesian classifier:

Procedure Bayesian classifier(x=<x_1,.....x_n>

//Document d is represented by x

begin

∀ classes $c_i \in c = c_1, \ldots \ldots c_m$

compute p(ci);

$\forall$ features $x_j \in x$

compute P($x_j|c_i$);

od;

multiply all P($x_j|c_i$) 's(=$\prod_{j=1}^{n} P(x_j|c_i)$);

calculate f_i(d)=P(c_i)*$\prod_{j=1}^{n} P(x_j|c_i)$;

od;

Assign d to the class(es)of max(f_1(d),.....$f_m(d)$

end.s

The advantages of naive bayes are: The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance. Three issues should be kept in mind, however. First, the naive Bayes classifier requires a very large number of records to obtain good results. Second, where a predictor category is not present in the training data, naive Bayes assumes that a new record with that category of the predictor has zero probability. This can be a problem if this rare predictor value is important. For example, consider the target variable bought high-value life insurance and the predictor category own yacht. If the training data have no records with owns yacht = 1, for any new records where owns yacht = 1, naive Bayes will assign a probability of 0 to the target variable bought high-value life insurance. With no training records with owns yacht = 1, of course, no data mining technique will be able to incorporate this potentially important variable into the classification model—it will be ignored. With naive Bayes, however, the absence of this predictor actively "out votes" any other information in the record to assign a 0 to the target value (when, in this case, it has a relatively good chance of being a 1). The presence of a large training set (and judicious binning of continuous variables, if required) helps mitigate this effect. The disadvantages are: A subtle issue ("disadvantage" if you like) with Naive-Bayes is that if you have no occurrences of a class label and a certain attribute value together (e.g. class="nice", shape="sphere") then the frequency-based probability estimate will be zero. Given Naive-Bayes' conditional independence assumption, when all the probabilities are multiplied you will get zero and this will affect the posterior probability estimate. This problem happens when we are drawing samples from a population and the drawn vectors are not fully representative of the population. Lagrange correction and other schemes have been proposed to avoid this undesirable situation.

C. Neural Network:NN is a non knowledge-based adaptive HDSS that uses a form of artificial intelligence, also known as machine learning, that allows the systems to learn from past experiences / examples and recognizes patterns in healthcare information. It consists of nodes called neuron and weighted connections that transmit signals between the neurons in a forward or looped fashion. It consists of 3 main

layers: Input which is data receiver, Output which communicates results or possible diseases and Hidden which processes data. The system becomes more efficient with known results for large amounts of data.[9]The advantages of NN include the elimination of needing to program the systems and providing input from experts. The NN HDSS can process incomplete data by making educated guesses about missing data and improves with every use due to its adaptive system learning. Additionally, NN systems do not require large databases to store outcome data with its associated probabilities. A neural network can perform tasks that a linear program can not. When an element of the neural network fails, it can continue without any problem by their parallel nature. Some of the disadvantages are that the training process may be time consuming leading users to not make use of the systems effectively. The NN systems derive their own formulas for weighting and combining data based on the statistical recognition patterns over time which may be difficult to interpret and doubt the system's reliability. The neural network needs training to operate. The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated. Examples include the diagnosis of appendicitis, back pain, myocardial infarction, psychiatric emergencies and skin disorders. The NN's diagnostic predictions of pulmonary embolisms were in some cases even better than physician's predictions.[10]

IV.RESEARCH METHODOLOGY

The proposed comparative analysis of considered data mining classification techniques have been done in WEKA. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA will be located in a folder called Weka-3.7 in the Program Files folder. The main program can be launched via a short cut or by clicking on a file called either weka.exe or weka.jar . Once launched, a small window will appear, usually in the top right of your screen.[13]

Fig 3 WEKA GUI Chooser

A. c4.5 algorithm implemented in WEKA

Here we are using the cardio vascular datasets, we are taking the datasets of heart datasets from UCI repository system and write it in notepad and save it in cardio vascular.arff format.[13]

@relation cleveland-14-cardio vasculart-disease

@attribute 'age' real

@attribute 'sex' { female, male}

@attribute 'cp' { typ_angina, asympt, non_anginal, atyp_angina}

@attribute 'trestbps' real

@attribute 'chol' real

@attribute 'fbs' { t, f}

@attribute 'restecg' { left_vent_hyper, normal, st_t_wave_abnormality}

@attribute 'thalach' real

@attribute 'exang' { no, yes}

@attribute 'oldpeak' real

@attribute 'slope' { up, flat, down}

@attribute 'ca' real

@attribute 'thal' { fixed_defect, normal, reversable_defect}

@attribute 'num' { '<50', '>50_1', '>50_2', '>50_3', '>50_4'}@data

Opening a data set.

In the Explorer window, click on "Open file" and then use the browser to navigate to the 'data' folder within the Weka-3.7 folder. Select the file called cardio vascular .arff.

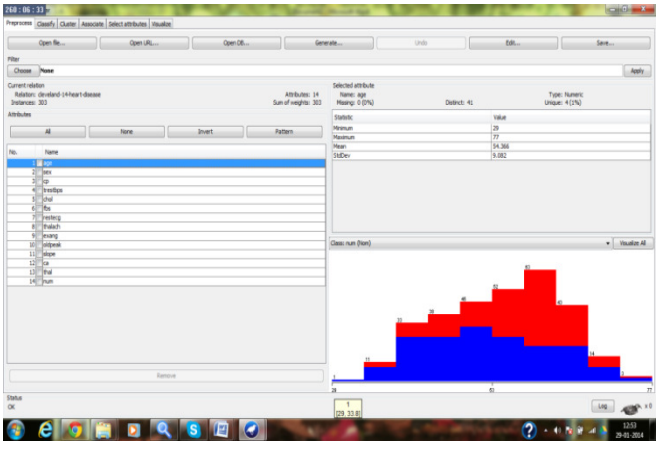

Fig 4 snapshot of opening datasets in WEKA explorer

Choosing a classifier

Next we must select a machine learning procedure to apply to this data. The task is classification so click on the 'classify' tab near the top of the Explorer window.

The window should now look like this:

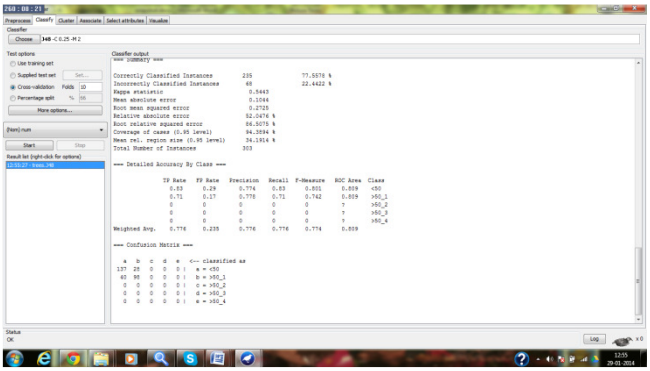

Fig 5 snapshot of classifier

V. RESULTS AND ANALYSIS

Both the data mining classification algorithms c4.5 and Bayesian Networks were implemented for the data set shown in section III and the results obtained were analyzed based on the following two parameters :

1. kappa statistic

Kappa statistic is a generic term for several similar measures of agreement used with categorical data . Typically it is used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories.

2. classified instances

correctly classified instances show the accuracy of the algorithm means +ve chances of the diseases in medical domain, Incorrectly classified instances show the accuracy of the algorithm means -ve chances of the diseases in medical domain.

Kappa Statistic: Kappa statistic is a generic term for several similar measures of agreement used with categorical data . Typically it is used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories. for example, kappa might be used to assess the extent to which (1) radiologist analysis of an x-ray, (2) computer analysis of the same x-ray, and (3) biopsy agree in labelling a growth "malignant" or "benign." Suppose each object in a group of M objects is assigned to one of n categories. The categories are at nominal scale. For each object, such assignments are done by k raters. The kappa measure of agreement is the ratio .

$$K = \frac{P(A) - P(E)}{1 - P(E)} \qquad (2)$$

where P(A) is the proportion of times the k raters agree, and P(E) is the proportion of times the k raters are expected to agree by chance alone. Complete agreement corresponds to $K = 1$, and lack of agreement (i.e. purely random coincidences of rates) corresponds to $K = 0$. A negative values of kappa would mean negative agreement - that is, the propensity of raters to avoid assignments made by other raters.

Review kappa scores and their implications:

0 No agreement beyond chance
0–0.2 Slight agreement beyond chance
0.2–0.4 Fair agreement beyond chance
0.4–0.6 Moderate agreement beyond chance
0.6–0.8 Substantial agreement beyond chance
0.8–1.0 Almost perfect agreement beyond chance

Examples of kappa scores for clinical correlation

In c4.5 algorithm: The kappa scores of different diseases are cardio vascular disease=0.5443,diabetes=0.4164,malignancy in lungs=0.3978

In Bayesian Networks: Here the kappa scores of different diseases are cardio vascular disease=0.6661,diabetes=0.429,malignancy in lungs=0.44

Graph of Kappa statistics

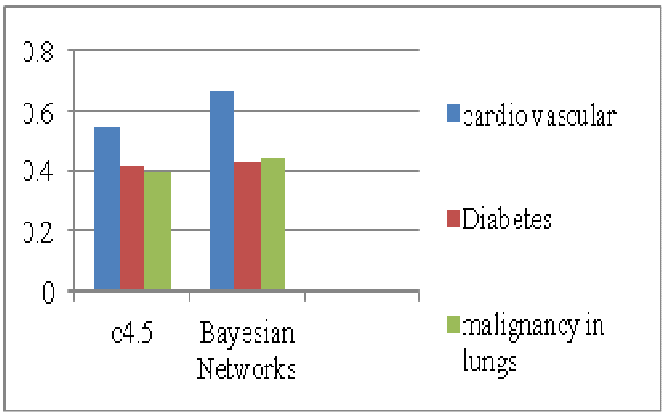

Fig 6 Kappa statistics

In this graph we found that the value of kappa statistics is high in cardio vascular disease. In Bayesian Networks the value of kappa statistics in cardio vascular, diabetes and malignancy in lungs is higher than c4.5.

Correctly and Incorrectly classified Instances in WEKA

Correctly classified Instances

	C4.5	Bayesian Networks
Cardio vascular	235	253
Diabetes	567	571
Malignancy in lungs	25	25

Table 1 correctly classified instances

Graph of correctly classified Instances

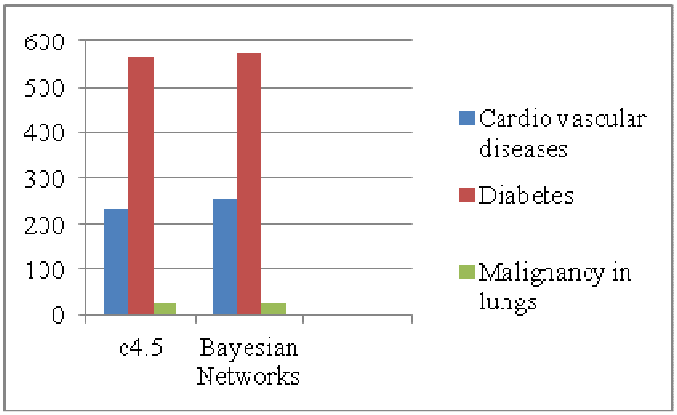

Fig 7 Graph of correctly classified instances in different algorithm in different disease

In this graph we found that the value of correctly classified instances in Bayesian Network is higher than c4.5.so Bayesian Network is better than c4.5

Incorrectly classified Instances

	C4.5	Bayesian Networks
Cardio vascular diseases	68	50
Diabetes	201	197
Malignancy in lungs	7	7

Table 3Incorrectly classified instances

Graph of Incorrectly classified Instances

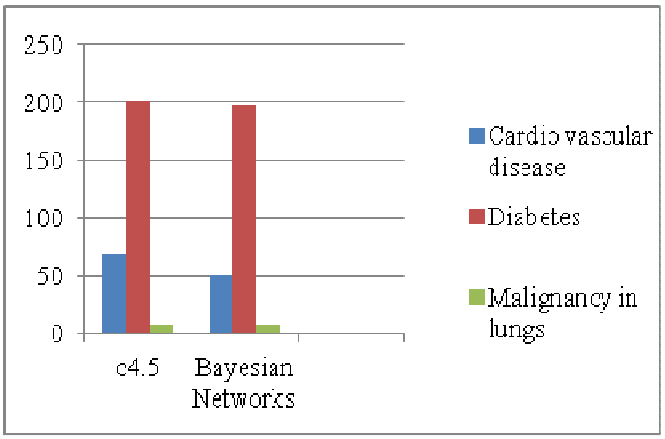

Fig 8 Graph of Incorrectly classified instances in different algorithm in different disease

In this graph we found that the value of Incorrectly classified instances in c4.5 is higher than Bayesian Networks. so c4.5 is lower than Bayesian Networks.

VI. DISCUSSION

Both the data mining classification algorithms c4.5 and Bayesian Networks were implemented for the data set shown in section III and the results obtained were analyzed based on the following two parameters

1.classified instances:correctly classified instances show the accuracy of the algorithm means +ve chances of the diseases in medical domain, Incorrectly classified instances show the accuracy of the algorithm means -ve chances of the diseases in medical domain.

2. kappa statistic:Kappa statistic is a generic term for several similar measures of agreement used with categorical data . Typically it is used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories.

In c4.5 algorithm: The kappa scores of different diseases are cardio vascular disease=0.5443,diabetes=0.4164,malignancy in lungs=0.3978

In Bayesian Networks: Here the kappa scores of different diseases are cardio vascular disease=0.6661,diabetes=0.429,malignancy in lungs=0.44

In Fig 6 we found that the value of kappa statistics is high in cardio vascular disease. In Bayesian Networks the value of kappa statistics in cardio vascular, diabetes and malignancy in lungs is higher than c4.5. In Fig 7 we found that the value of correctly classified instances in Bayesian is higher than c4.5.so Bayesian Network is better than c4.5. In Fig 8 we found that the value of Incorrectly classified instances in c4.5 is higher than Bayesian Networks. so c4.5 is lower than Bayesian Networks. So finally we found that Bayesian Network is much more better than c4.5.

VII. CONCLUSION

The main goal of this paper is to analyze and implement the data mining algorithms using weka tool and comparison between c4.5 and Bayesian Networks**.** Two algorithms were chosen: c4.5, and Bayesian Networks.+ve tested the result of Diabetes is very high. Here the values of Bayesian Network is greater than c4.5. that in -ve tested the result of Diabetes is high. Here the values of c4.5 is greater than Bayesian Networks. Our further future work for this paper is to implement other algorithms like neural network and clustering with use of medical datasets in WEKA.

REFERENCES

[1]. Mariscal, Gonzalo, Óscar Marbán, and Covadonga Fernández. "A survey of data mining and knowledge discovery process models and methodologies."Knowledge Engineering Review 25.2 (2010): 137.

[2]. Lokanatha C. Reddy, A Review on Data mining from the Past to the Future, International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011
[3]. Varsha Kavi and Divyesh Joshi , "A Survey on Enhancing Data Processing of Positive and Negative Association Rule Mining", International Journal of Computer Sciences and Engineering, Volume-02, Issue-03, Page No (139-143), Mar -2014
[4]. Ozer, Patrick. "Data Mining Algorithms for Classification." (2009).
[5]. Drazin, Sam, and Matt Montag. "Decision Tree Analysis using Weka." Machine Learning-Project II, University of Miami (2012): 1-3.
[6]. Drazin, S., & Montag, M. (2012). Decision Tree Analysis using Weka. Machine Learning-Project II, University of Miami, 1-3.
[7]. Bouckaert, Remco R. "Bayesian network classifiers in weka for version 3-5-7." Artificial Intelligence Tools 11.3 (2008): 369-387.
[8]. Bouckaert, Remco R. Bayesian network classifiers in weka. Department of Computer Science, University of Waikato, 2004.
[9]. Singh, Yashpal, and Alok Singh Chauhan. "Neural networks in data mining." Journal of Theoretical and Applied Information Technology 5.6 (2009): 36-42.
[10]. Suyal, Neha. "Data Mining Using Neural Networks."
[11]. bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006
[12]. Sharma, Trilok Chand, and Manoj Jain. "WEKA Approach for Comparative Study of Classification Algorithm."
[13]. Arbelaitz,Olatz, etal. "J48Consolidated: An implementation of CTC algorithm for WEKA¥." (2013).

