Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis

Prof. Dhomse Kanchan B.
Assistant Professor of IT department MET'S
BKC IOE, Nasik Nasik, India
kdhomse@gmail.com

Mr. Mahale Kishor M.
Technical Assistant of IT department
MET'S BKC IOE, Nasik, India
kishu2006.kishor@gmail.com

***Abstract* – The worldwide study on causes of death due to heart disease/syndrome has been observed that it is the major cause of death. If recent trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030. The healthcare industry collects large amounts of heart disease data which unfortunately are not "mined" to discover hidden information for effective decision making. In this paper, study of PCA has been done which finds the minimum number of attributes required to enhance the precision of various supervised machine learning algorithms. The purpose of this research is to study supervised machine learning algorithms to predict heart disease. Data mining has number of important techniques like categorization, preprocessing. Diabetic is a life threatening disease which prevent in several urbanized as well as emergent countries like India. The data categorization is diabetic patients datasets which is developed by collecting data from hospital repository consists of 1865 instances with dissimilar attributes. The examples in the dataset are two categories of blood tests, urine tests. In this research paper we discuss a variety of algorithm approaches of data mining that have been utilized for diabetic disease prediction. Data mining is a well known practice used by health organizations for classification of diseases such as diabetes and cancer in bioinformatics research.**

Keywords – Support, Vector Machine, Naïve Bayes, Decision Tree, Principal Component Analysis, Diabetic Disease Prediction.

I INTRODUCTION

A.Data mining:

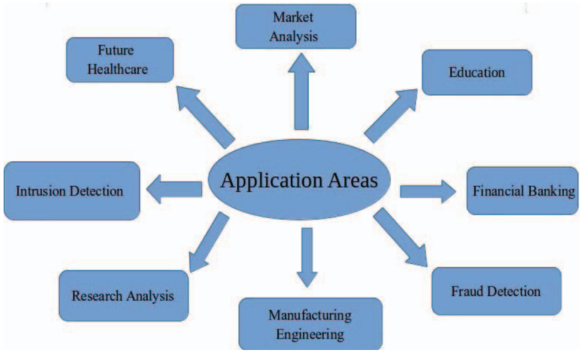
The process of identifying commercially useful patterns or relationship in databases or other Computer repositories through the use of advanced statistical tools are known as Data Mining. It is a relatively new and promising technology. Data can be analyzing from different perspectives and summarizing it into useful information- information which can be used to discovering significant new correlation, patterns and trends by digging into large amounts of data stored in warehouse; using statistical, machine learning.

AI: An artificial intelligence & data visualization techniques. Industries like medical, manufacturing, aerospace, chemical etc that are already taking benefit of data mining. The expertise usually consent that in-depth decision support requires new technology. This new technology should enable the innovation of trends and predictive patterns in data the creation and testing of hypothesis and generation of insight-provoking visualizations. The concept of data mining helps the end users to extract useful information from large databases. These large databases are present in data warehouses, i.e., "Data Mountain," which are presented to data mining tools. In short data warehousing allows one to construct the data mountain. Data mining is the nontrivial removal of implicit, previously unidentified and potentially useful information from the data mountain. This data mining is not precise to any industry – it requires intellectual technologies and the eagerness to explore the opportunity of hidden knowledge that resides in the data. Also data mining is known as knowledge discovery in databases (KDD). Data mining is concerned to finding hidden relationships present in business data to allow businesses to make predictions for Future use. It is the process of datadriven removal of not so clear but useful information from large databases.

1. Machine Learning:

We are entering era of big data. For example, there are about 1 trillion web pages; 1 hour of video is uploaded to YouTube each second amounting to ten years of content every day; the genomes of thousands of people each of which has a length of $3.8x10^9$ base pairs have been sequences by a variety of labs. Wal-Mart handles more than 1Milion transactions every hour and has databases containing larger than 2.5 x 10^{15} of information Cukier 2010 and so on. So this overflow of data calls for preset methods of data analysis which is what machine learning provides. We define ML as a set of methods that can automatically identify patterns in data and then use the uncovered patterns to forecast future data or to perform other kinds of decision making under uncertainty. ML is a set of tools that rarely speaking allows us to "teach" computers how to perform tasks by providing the examples of how they should be done.

Fig.1: Applications of Data Mining



2. Types of Machine Learning

Some of the main types of machine learning are:

a) Supervised Learning:

In this learning the training data is labeled with the correct answers for example spam or ham the two most common types of supervised learning are classification and regression.

b) Unsupervised learning:

In this we are given a collection of unlabeled data, which we wish to analyze & discover patterns within. The two most important examples are dimension reduction and clustering.

c) Reinforcement learning:

In this an agent for example, a robot or controller seeks to learn the optimal actions to take based the results of past actions. There are various other types of ML for example:

1. Semi-supervised learning, in which only a subset of the training data is labeled in the system.

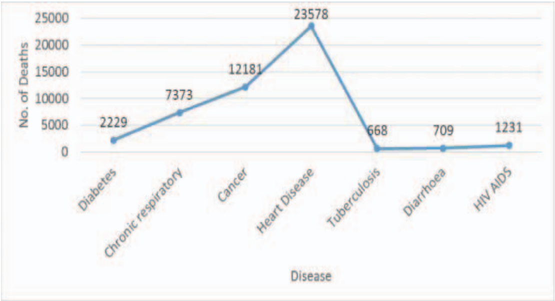
2. Time-series forecasting, such as in financial markets in the system.

3. Anomaly detection like used for fault-detection in factories and in surveillance process.

4. Active learning which obtaining the data is more costly and so an algorithm must establish which training data to obtain and many others.

Based on a recent study by the Registrar General of India or RGI and the Indian Council of Medical Research in the age group of 25-69 years about 25 % of deaths occur because of heart diseases [1]. It is the single largest cause of death in the globe. Several researchers are using statistical & data mining tools in the diagnosis of heart disease. There are various complex data mining techniques and algorithms which are used in various areas [2] for forecast. Some of the Application areas of data mining are given in Figure 2. Data mining is an essential step of knowledge discovery. It combines statistical analysis, ML & database technology to extract hidden patterns and relationships from databases. Data mining uses two strategies: 1) supervised learning and 2) Unsupervised learning.

Fig. 2: Projected number of deaths worldwide by 2030



DM shows a method developed to inspect huge amounts of data regularly gathered. Generally the term also refers to a collection of tools used to perform the process. One of the constructive applications in the field of medicine is the incurable chronic disease diabetes. DM algorithm is used for verifying the accuracy in predicting diabetic status. Classification is one of the most repeatedly studied problems by DM and ML researchers. Classification consists of predicting a certain results based on a given input. In order to predict the result the algorithm processes a training set including a set of attributes and the respective results generally called goal or prediction attribute in the system. Here the algorithm tries to determine relationships between the attributes that would make it possible to predict the results. Then the algorithm is given a data set not seen before called training set which includes the same set of attributes instead of the class label not yet known. Here, the algorithm determines the input and gives a prediction in the process.

III IMPLEMENTATION

A. Naive Bayes classification:

The Bayesian Classification represents a supervised learning method also a numerical method for classification. It assumes an essential probabilistic model and it permit us to confine uncertainty about the model in an ethical way by determining probabilities of the results. So it can solve diagnostic and predictive problems of the process. In this Classification the named after Thomas Bayes, 1702-1761 who proposed the Bayes Theorem; the Bayesian classification gives practical learning algorithms, prior knowledge and observed data can be collected.

a) Mathematical Representation of Naive Bayes classification

i) Let T: Set of tuples

Each Tuple is an ‘n’ dimensional vector

S: $(s_1,s_2,s_3,s_4,\ldots\ldots s_n)$

Where s_i is the value of attribute A_i

Let there are ‘m’ Classes: $C_1, C_2, C_3, C_4\ldots C_m$

Bayesian classifier predicts S belongs to Class C_i

ii)Maximum Posteriori Hypothesis

$$\text{Posterior Probability} = \frac{\text{Prior x likelihood}}{\text{Evidence}}$$

$$P(C_i|S) = \frac{P(S|C_i)\,P(C_i)}{P(S)}$$

The Bayesian Classification gives a valuable perspective for understanding and examines various learning algorithms. So, it calculates precise probabilities for hypothesis and it is robust to noise in input data in the classification.

B. Heart Disease Data

For the study of heart disease data the researcher use a dataset from Cleveland Clinic Foundation [3]. In this dataset have 11 attributes/characters and 303 rows/records. Every row corresponds to one specific patient and every attribute corresponds to the observations or type of tests for patients/substance. The explanation of the attributes is shown in Table 1

Table 1: Attributes and their description

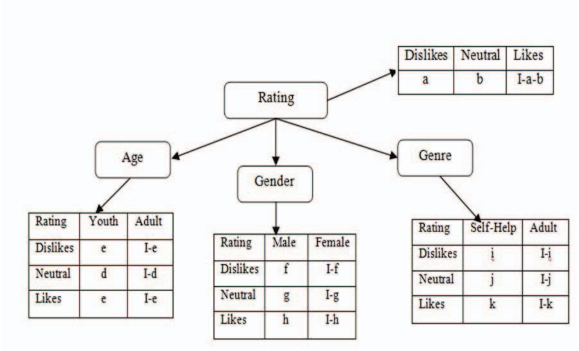
No.	**Attributes**	**Description**
1	Age	Age in years
2	Gender	1 Male 0 Female
3	Chest Pain Type	1 Typical Angina 2 Atypical Angina 3 Non-Angina pain 4 Asymptomatic Pain
4	Blood Pressure	Blood pressure in mm Hg
5	Cholesterol	Cholesterol level in mg/dl
6	Blood Sugar	Is blood Sugar > 120 mg/dl 1 True 2 False
7	Resting ECG	0 Normal 1 Having ST-T wave abnormality 2 Showing probable or define left ventricular hypertrophy
8	Max Heart Rate	Maximum Heart Rate Achieved

C. Algorithms

a) Naive Bayes classifier

The supervised machine learning method of classification is represented by Naive Bayes algorithm. It uses a probabilistic model by determining probabilities of the outcomes/outputs. It is used in analytical and predictive problems. Naive Bayes is robust to noise in input dataset. An implementation of Naive Bayes has been illustrated in Figure 3.

Fig.3: Naive Bayes



Dislikes	Neutral	Likes
a	b	I-a-b

Rating	Youth	Adult
Dislikes	e	I-e
Neutral	d	I-d
Likes	e	I-e

Rating	Male	Female
Dislikes	f	I-f
Neutral	g	I-g
Likes	h	I-h

Rating	Self-Help	Adult
Dislikes	i	I-i
Neutral	j	I-j
Likes	k	I-k

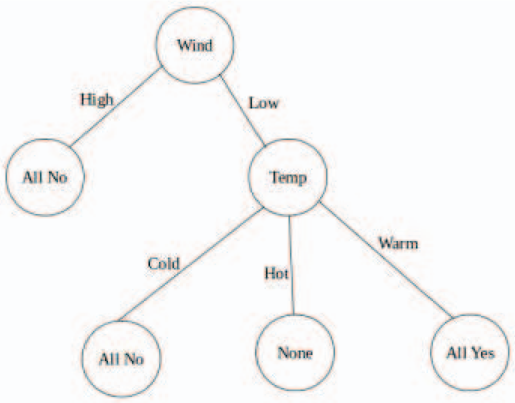
b) Decision Tree

The decision tree learning is like as decision tree algorithm which uses maps input about an item to output of the item. The tree models with finite classes of output are called classification trees. In these tree structures leaves shows class labels and branches shows relation between attributes that the results in those class labels of the system. Decision trees with continuous output classes are called regression trees. In data mining, a decision tree can be an input for decision making. An example of decision tree is demonstrated in Figure 4.

c) Support Vector Machine

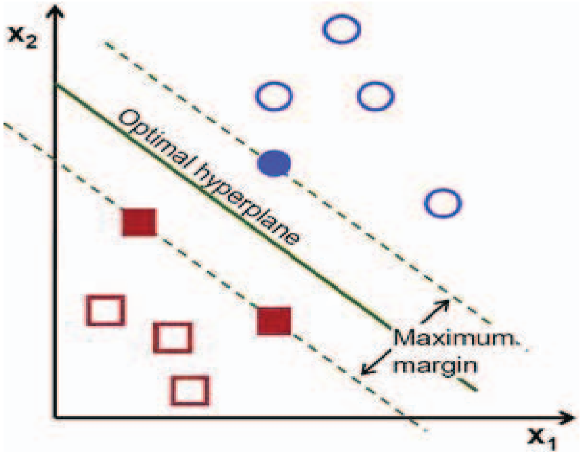
The SVM is a supervised machine learning algorithm for margin classification. It puts a hyper plane between the classes. SVM performs classification tasks by maximizing the margin which separates the classes while minimizing the classification errors.

Fig 4: Decision Tree



The working of Support Vector Machine is given in Figure 5.

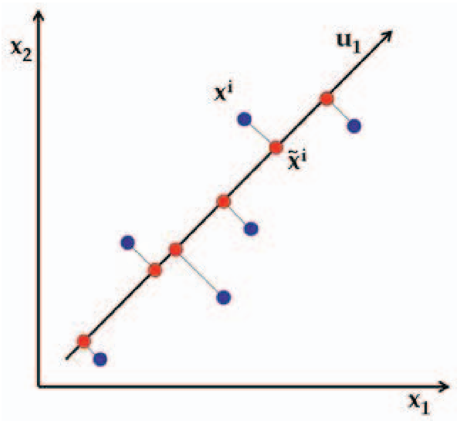
Fig.5: Support Vector Machine



d)Principal Component Analysis

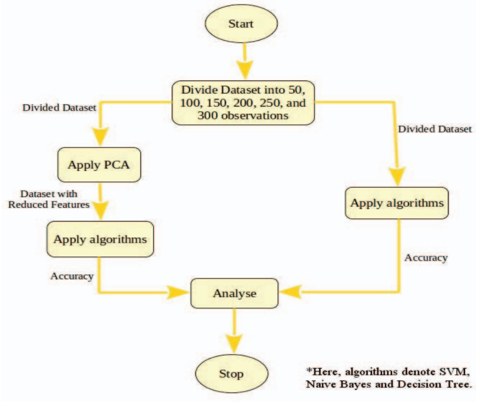
PCA is a arithmetic procedure that uses an orthogonal conversion to convert a set of observations of possibly allied variables into a set of standards of linearly uncorrelated variables called principal components in our system. Here, the initial principal component has the biggest possible variance [5]. So, the variance decreases after each consequent principal component in the system. Here, the resulting vectors are an uncorrelated basis set 31. The working of PCA is given in Figure 6.First the dataset is divided into training dataset and test dataset. The training dataset is being fed into the algorithms. The algorithms learn from this dataset. Later, in the test dataset, all the columns except the last one are fed in the algorithms. The last column is the actual outcome. The algorithm with the input data forms a column of its own. It can do so because it has learned the pattern from the training dataset.

Fig.6: Principal Component Analysis



The predicted column given by the algorithm is then compared to the actual column in the dataset. This comparison gives the required accuracy. The work-flow of this research work has been depicted as a flowchart in Figure 7.

Fig.7: Flowchart of the approach


Start
Divide Dataset into 50, 100, 150, 200, 250, and 300 observations
Divided Dataset
Divided Dataset
Apply PCA
Dataset with Reduced Features
Apply algorithms
Apply algorithms
Accuracy
Accuracy
Analyse
Stop
*Here, algorithms denote SVM, Naive Bayes and Decision Tree.

B. Diabetes Disease Data:

Explorer, Experimenter and Knowledge flow are the interface existing in WEKA that has been used by the system. Here we have used these data mining techniques to predict the chances of survivability of Diabetes disease through classification of different algorithms accuracy [8- 10].

C. Classification

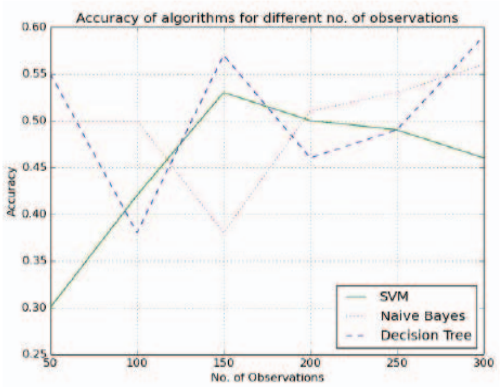
In data mining tools classifications deals with identifying the problem by observing characteristics of diseases along with patients and diagnoses or forecast which algorithm shows finest performance on the basis of WEKA's statistical results. Here three methods have been acquired in the paper. Here, the initial technique uses explorer interface and depends on algorithms such as Naïve Bayes, which is used in areas to represent and utilize and learn the statistical knowledge and significant outcomes have been gained [6]. Then the second technique uses Experimenter interface and this study permits one to design experiments for current algorithms like Naïve Bayes on datasets [7]. The third technique utilizes Knowledge Flow. Here we categorized the correctness of different algorithms Naïve Bayes, on different data sets.

IV RESULT ANALYSIS:

A. For Heart Disease:

In this work, the dataset is fed in the intervals of 50. The sequences in which the observations are given are 50, 100, 150, 200, 250 and 303 observations respectively. These observations are then divided into train and test dataset. Both datasets are then combined with three classifiers to predict values. Based on their performance, their accuracy is generated. When the accuracies of algorithms are plotted against the number of observations, a graph is generated. The graph is shown in Figure 8.

Fig.8: Number of observations vs Accuracy



B. For Diabetes Disease:

The data mining techniques that have been used by us using algorithms Naïve Bayes Through these techniques we trained out results on the basis of time taken to build model, correctly classified instances, error and ROC area. Algorithm scoring accuracy is shown in Naïve Bayes 34.8958 % correctly instances accuracy with minimum Naïve Bayes Mean Absolute Error

= 0.2841 having maximum Naïve Bayes ROC =0.819 time

taken to build model=0.02 seconds Classification Accuracy=34.8958 So from Explorer Interface data mining technique we can deduce that Naïve Bayes have maximum accuracy, least error and it takes less time to build model it and has maximum ROC.

Fig.9: Accuracy of SVM Before and After Feature Selection

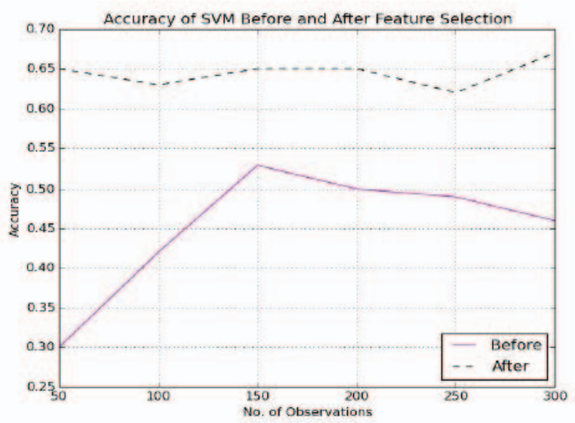


Fig.10: Accuracy of Naive Bayes before and After Feature Selection

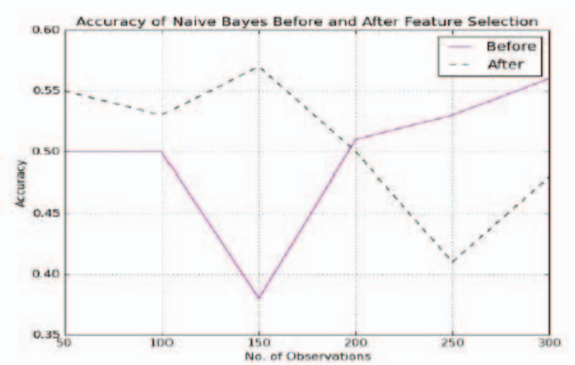
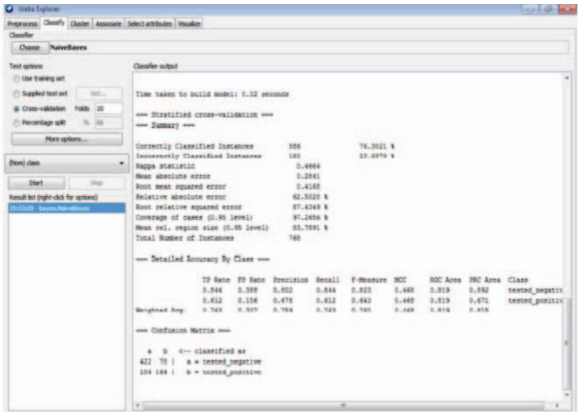


Fig.11. Naïve Bayes Algorithm applied on the Diabetic dataset



Classifier: NaiveBayes

Cross-validation Folds: 10

Percentage split %: 66

(Nom) class

19:53:00 - bayes.NaiveBayes

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Coverage of cases (0.95 level)          97.2656 %
Mean rel. region size (0.95 level)      83.7891 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tested_negativ
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tested_positiv
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

CONCLUSION

In this paper, for heart disease prediction SVM, Naive Bayes and Decision tree has been applied with and without using PCA on the dataset. We used PCA to reduce the number of attributes. After reducing the size of the dataset, SVM outperforms Naive Bayes and Decision tree. SVM can further be used to predict heart disease. A GUI desktop application can be built using SVM and this dataset to predict the possibility of cardiovascular disease in a patient and for diabetes data prediction, the main aim of this paper is to predict diabetes disease using WEKA data mining tool. Our algorithms were implemented using WEKA data mining technique to analyze algorithm accuracy which was obtained after running these algorithms in the output window. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, time taken to build model, mean absolute error and ROC Area. So, using above all observations, we can conclude that Maximum ROC Area means excellent predictions performance as compared to other algorithms.

ACKNOWLEDGEMENT

It's an honor to work with my IT department and our Institute and thankful for the great support regarding paper and their all views. We are really thankful to my all staffs and our H.O.D. Prof. N.R. Kale who give me the way of successful journey of making the paper and developing this Concept.

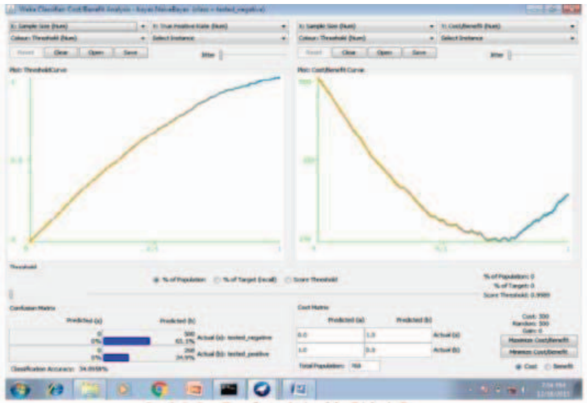


Fig.12. Cost/Benefit Analysis of the Diabetic Dataset

REFERENCES

[1] Vikas Chaurasia, Saurabh Pal, - "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.J.SciTech, 2013, Vol. 1, 208-2017

[2] http://bigdata-madesimple.com/14-useful-applications-of-data-mining

[3] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[5] M. Pechenizkiy, A. Tsymbal and S. Puuronen, "PCA-based feature transformation for classification: issues in medical diagnostics", IEEE, Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings (1063- 7125), Page No. 535 – 540, June 2004

[6] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 4, 2013, Page 56-66.

[7] S , Liver Disease Prediction Using Bayesian Classification , Special Issues , 4th National Conference on Advance Computing , Application Technologies, May 2014

[8] SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.

[9] Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.

[10] David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38, 2013