COMPARING VARIOUS CLASSIFIER TECHNIQUES FOR EFFICIENT MINING OF DATA

Harsha Jain[1], Dheeraj Pal[2], Alok Jain[3]

[1,2,3]Computer Science Engineering, Amity University (M.P) Gwalior, (India)

ABSTRACT

With recent advances in computer technology large amounts of data could be collected and stored. But all this data becomes more useful when it is analyzed and some dependencies and correlations are detected. This can be accomplished with machine learning algorithms. WEKA (Waikato environment for knowledge analysis) is a collection of machine learning algorithms implemented in Java. WEKA consists of a large number of learning schemes for classification and regression numeric prediction. So by using this we can find out the prediction value of dataset and the data which we stored can be seen in different forms in the form of matrix, graph, curve, tree etc. In this paper we are researching or comparing the results of the three classifiers, the classifiers we are using such as J48, Naïve Bayes, and pre-process the data. We compare the results which provide easy way to understand all the datasets and its condition.

Keywords:* *Data mining, Classification, Clustering, Visualization, Weka, Data-preprocessing, Algorithms

I. INTRODUCTION

The changes that are constantly taking place in terms of rapid technical and technological developments affect society as a whole. Companies invested in building data ware house that contains millions of records & attributes. They cannot produce sufficient output due to lack of knowledge, lack of staffs & appropriate tools. In recent years, there has been increasing interest in the use of data mining **[6], [1]** to investigate scientific questions with in the details of company employees. An area of enquiry about the details of employees, which is influenced by many qualitative, attributes such as income of employee, age, marital status, sex, education **[4],[13]**. Over all by this information we can predict that how many of the employees get the minimum salary, maximum salary, mean and the standard deviation for the salary, and we can calculate how many of employees are single, married, divorced, and their education like highschool, graduation, college etc.For this we are applying many techniques to predict how many of get how much salary and their overall status.

II. LITERATURE SURVEY

Data mining, as the confluence of multiple disciplines, including machine learning , statistics, database systems, information science, visualization, and many application domains, has made great progress in the past decade**[13].** Though many methods are proposed to address the issue of imbalanced classification, but still the solutions are problem dependent **[7].**As the previous papers using techniques rapid miner tool, orange, and weka but most of the results are come out from rapid miner tool and orange and by using these tools they apply technique such as k-nearest neighbour, Neural network, Fuzzy, Genetic and other techniques are applied in the environment **[4].** This paper describes the comparative performance of classification techniques **[4].** H.2.8 [Database Management]: Database Applications—Data Mining, Spatial databases and GIS; H.5.1 [Information Interfaces] in the paper **[5]** for finding the time series data.

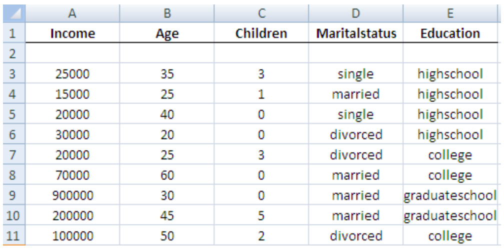
III. STUDY

3.1 Hypothesis

The research objective was to find a very large amount of data which is collected from the local serve areas and run data mining algorithms against it. Popular data mining software is used to evaluate the data, tool is weka **[12]** which provides a superior result **[7].**Additionally the raw data was graphed to visualize **[6]** the identified patterns through visual inspection which the mining software might overlook. By mining the data a trend was expected about the employee's information how many of come under maximum, minimum, mean and standard deviation.

3.2 Data Collection and Preparation

The dataset **[4], [13]** was stored in a table, each type of data given an own columns in a table with in a database. From Table.1, represent the qualitative value which is gathered for performing the experiment. All types of data are stored in a same table so it is easy for the software to calculate the data.**Table.1** shows the attributes there are five attributes income, age, children, maritalstaus, education and total 10 instances.



	A	B	C	D	E
1	Income	Age	Children	Maritalstatus	Education
2					
3	25000	35	3	single	highschool
4	15000	25	1	married	highschool
5	20000	40	0	single	highschool
6	30000	20	0	divorced	highschool
7	20000	25	3	divorced	college
8	70000	60	0	married	college
9	900000	30	0	married	graduateschool
10	200000	45	5	married	graduateschool
11	100000	50	2	divorced	college

Table.1 A Data which is collected for Experiment.

The data cannot be directly compared to each other because they have different form of values such as income is in numeric form and marital status was in nominal **[3], [10].** So here we are only calculating the maximum and minimum form. With the help of **TABLE.2** it can define the attributes and the values, for predicting the data.

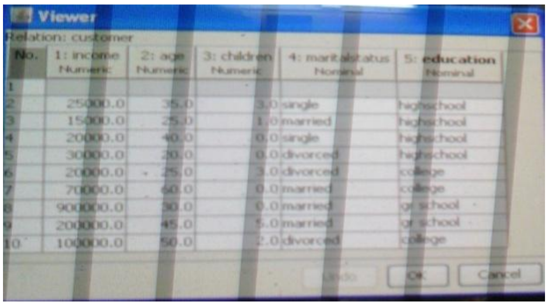
Attributes	Values
Income	15000 to 900000
Age	20-50
Marital status	Single,Married,Divorced
children	0-5
Education	highschool,grad.school,college

TABLE.2 Collection of qualitative data.

3.3 Data Pre-Processing

Data shown in **TABLE.1** is processed in the software; it would be quite massive to remake the table for local database. So a java program was used which can pull data from the database and convert it in to data mining format **[10]**. The Program then compiles the result in to a weka file form. The file can be read like a table which has its own column. This lets data be efficiently organized and allow for mining techniques.

TABLE.2 Below the table was readable by software directly.



Viewer

Relation: customer

No.	1: income Numeric	2: age Numeric	3: children Numeric	4: maritalstatus Nominal	5: education Nominal
1					
2	25000.0	35.0	3.0	single	highschool
3	15000.0	25.0	1.0	married	highschool
4	20000.0	40.0	0.0	single	highschool
5	30000.0	20.0	0.0	divorced	highschool
6	20000.0	25.0	3.0	divorced	college
7	70000.0	60.0	0.0	married	college
8	900000.0	30.0	0.0	married	gr school
9	200000.0	45.0	5.0	married	gr school
10	100000.0	50.0	2.0	divorced	college

Now, with the help of this table we can easily select the attributes as it is save in the software i.e. in the "WEKA" tool.

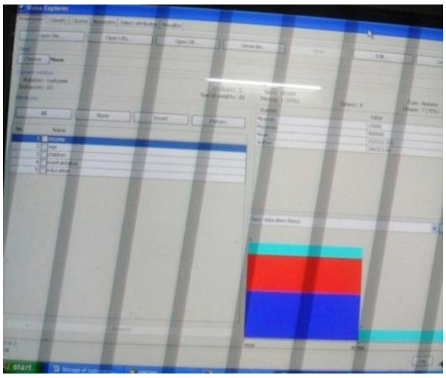


Fig: 1 Preprocessing Qualitative Data.

It can be noted from the related work that the attribute selection plays an important role, to identify parameters that are important and significant for an excellent result. Where Blue, Red, skyblue colour represents the no. of maximum, minimum, mean values. Blue is for maximum, Red is for mean and skyblue is for minimum by which it is easy to identify that their income, age, children, maritalstaus, education lies between the given data. In **Fig.2** we can clearly see all the dataset at a time by selecting the option visualize all the data or all the attributes**.[6],[5],[9].**

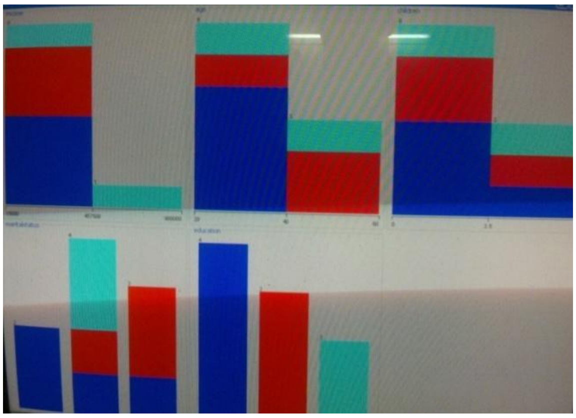


Fig: 2 Visualize the overall Dataset.

IV. METHODOLOGY OF DATA MINING TECHNIQUES FOR CLASSIFICATION

4.1. Classification Techniques for Prediction

In an effort to finding patterns, a variety of algorithms are used **[10], [11].** There are three main algorithms that provide some form of result, and those were the:

- Classification Rules Algorithm
- Various Decision Tree Algorithms
- Naive Bayes.

Many classification **[8], [10]** models have been proposed by researchers in Machine learning, pattern recognition and in statics. Generally the classification techniques can follow the two steps process which is used to predict the class labels for training data. In classification step, test data are used to estimate the accuracy of classification rules. There are many techniques that can be used for classification techniques. Such as techniques are Decision tree, NayeBayes rule any many other**. Fig: 1** Represents the classification methodology of research process can be learned from training data which are analyzing from classification algorithm. Test data are used to classification estimate the accuracy of classification rules.

4.2 Flow Chart

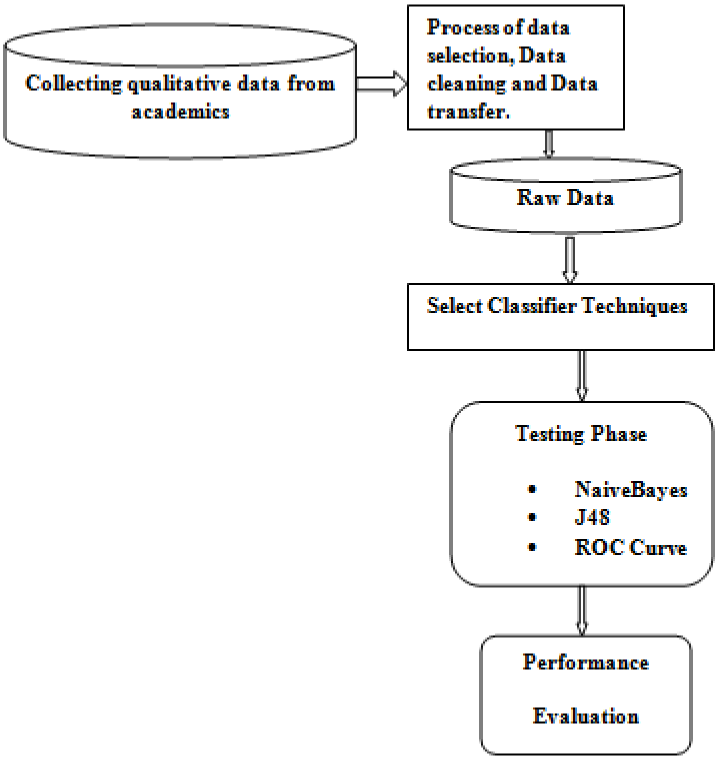

Collecting qualitative data from academics
Process of data selection, Data cleaning and Data transfer.
Raw Data
Select Classifier Techniques
Testing Phase
• NaiveBayes
• J48
• ROC Curve
Performance Evaluation

Fig: 1 Methodology of Classification Technique used in Data Mining

4.3 J48.Tree Classifiaction Technique

J48 is a tree technique which is enhanced by ID3 algorithm. It is one of the most popular tree algorithm in tree classification technique with the help of this tree we can determine the number of employees between in which criteria. The objective is to reduce the impurity or uncertainty in data as much as possible.

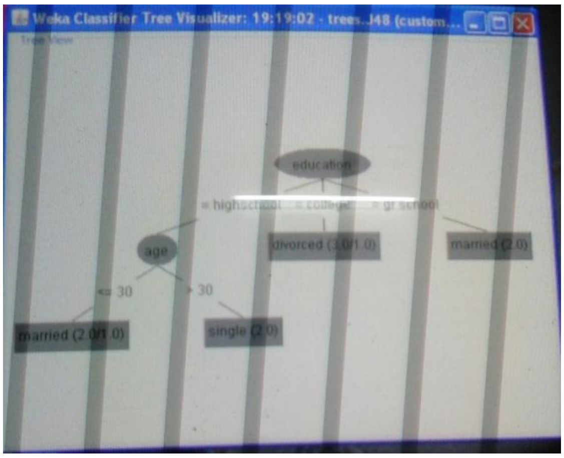


Fig: 3 "J48 the Decision Tree" has been Built

Decision tree is very effective in mining the data with the help of this tool it is easy to reduce the bulk of data, and clearly understood the details or the information with the help of this tree.

This algorithm is an extension of ID3 algorithm and possibly create small tree. It uses a divide and conquers approach. Decision tree is closely related to the rule indication. A tree includes root node i.e. "education "and internal node that represents test condition (applied on attributes) a leaf node"married, single, divorced".

4.4 Naive Bayes Classifier

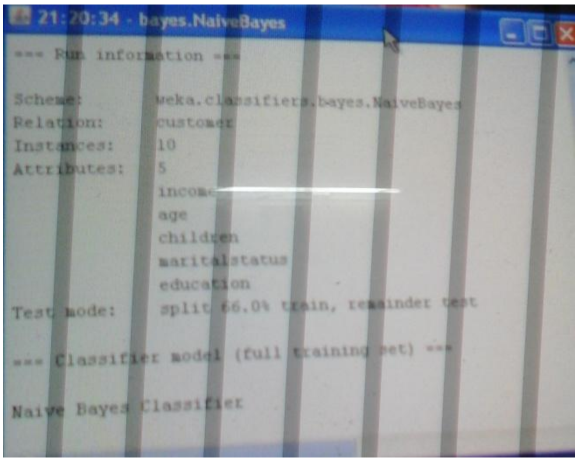
In probability theory, Naive Bayes classifier checking the condition rule and it can be classified by learning phase and testing phase. Bayesian reasoning is applied to decision making that deal with probability inference which is used to gather the knowledge of prior events by predicting events through rule base. Once the model has been trained and tested, we need to measure the performance of the model .For this purpose we use three measures namely: precision, recall and accuracy.

Precision (P) = tp/(tp+fp)

Recall (R) = tp/(tp+fn)

Accuracy (A) = (tp+tn)/Total

Where tp, fp, tn and fn are true positive, false positive, true negative and false negative respectively.

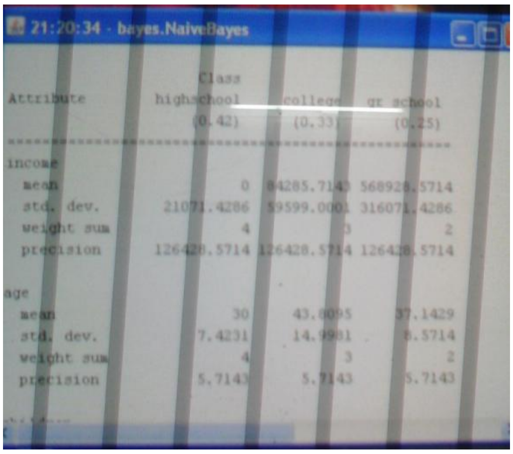


```
21:20:34 - bayes.NaiveBayes

=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     customer
Instances:    10
Attributes:   5
              income
              age
              children
              maritalstatus
              education
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

Naive Bayes Classifier
```

Fig: 4 "Naïve-Bayes Classifiers"

Above **fig: 4** show the running information with the help of bayes classifiers.

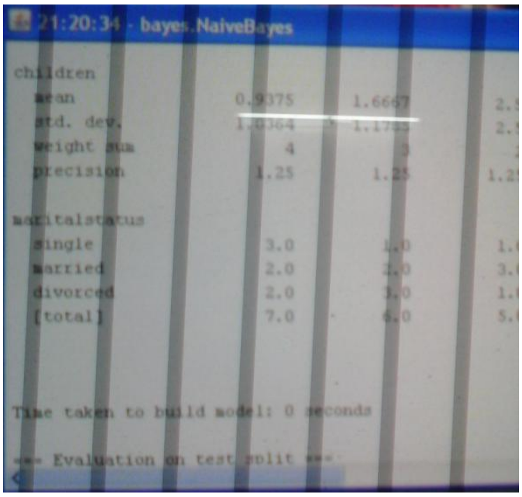
That while performing Naïve bayes in this Test Mode is used percentage Split (split 66.0% default value).



```
21:20:34 - bayes.NaiveBayes

                              Class
Attribute                highschool    college   gr school
                             (0.42)     (0.33)      (0.25)
==========================================================
income
  mean                            0 84285.7143 568928.5714
  std. dev.              21071.4286 59599.0001 316071.4286
  weight sum                      4          3           2
  precision             126428.5714 126428.5714 126428.5714

age
  mean                           30    43.8095     37.1429
  std. dev.                  7.4231    14.9981      8.5714
  weight sum                      4          3           2
  precision                  5.7143     5.7143      5.7143
```

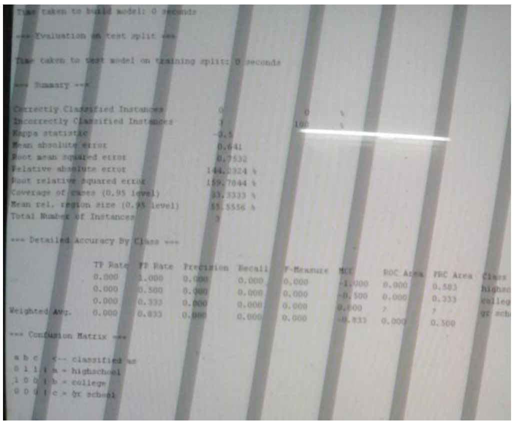
Fig: 5 NaiveBayes shows “values mean maximum, minimum, standard deviation.

Naïve bayes shows the attributes income, age, children, maritalstatus and the class is education its attributes is high school, college, graduation. With the help of this technique mean, std.deviation, weight, precision is calculated for each and every attribute.



```
21:20:34 - bayes.NaiveBayes

children
  mean                   0.9375     1.6667     2.5
  std. dev.              1.0364     1.1785     2.5
  weight sum                  4          3       2
  precision                1.25       1.25     1.2

maritalstatus
  single                    3.0        1.0     1.
  married                   2.0        2.0     3.
  divorced                  2.0        3.0     1.
  [total]                   7.0        6.0     5.



Time taken to build model: 0 seconds

=== Evaluation on test split ===
```

Fig: 6 “Naïve Bayes Running Image”



```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances         3             100      %
Kappa statistic                         -0.5
Mean absolute error                      0.641
Root mean squared error                  0.7532
Relative absolute error                144.2324 %
Root relative squared error            159.7044 %
Coverage of cases (0.95 level)          33.3333 %
Mean rel. region size (0.95 level)      55.5556 %
Total Number of Instances                3

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
               0.000    1.000    0.000      0.000   0.000      -1.000  0.000     0.583     highsc
               0.000    0.500    0.000      0.000   0.000      -0.500  0.000     0.333     colleg
               0.000    0.333    0.000      0.000   0.000      0.000   ?         ?         gr sch
Weighted Avg.  0.000    0.833    0.000      0.000   0.000      -0.833  0.000     0.500

=== Confusion Matrix ===

 a b c   <-- classified as
 0 1 1 | a = highschool
 1 0 0 | b = college
 0 0 0 | c = gr school
```

Fig: 7 Shows the “Confusion Matrix.”

Below the table shows that what proportion of test instances has been correctly and incorrectly classified.

Correctly Classified Instances 8 88.8889%
Incorrectly classified Instances 1 11.1111%

Table.1 Shows the accuracy of Instances

This is the correct accuracy rate according to the no. of instances. At the bottom **Table.2** shows a "confusion matrix"

a	b	c	Classified as
4	0	0	a = high school
1	2	0	b= college
0	0	0	c= Graduate school

Table.2 Confusion Matrix

4.5 Reciever Operating characteristics (ROC) Curve

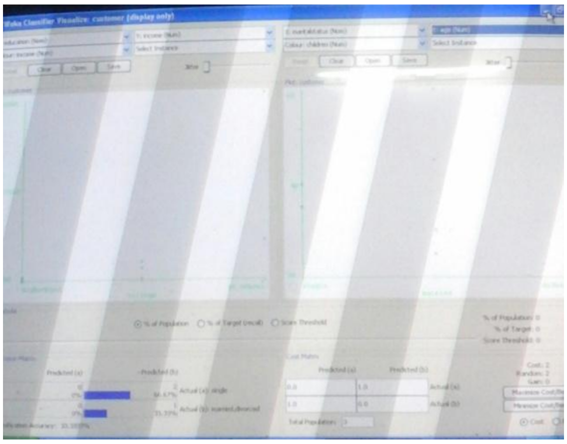


Fig: 8 "ROC curve of J48 Tree"

This curve can be described as follows there are two curve can be seen together with different attributes and class and it is easy to find out the maximum and minimum instances of the values. In **table.1** represents first graph its X-axis shows education (nominal values) in which highschool, school and graduation, and its Y-axis shows shows income (numeric values) while in next **table.2** graph its X-axis represents the marital status (nominal value) Y-axis represents the age (numeric value).And the jitter of graph first shows education of the candidates. While the jitter of graph two represents childrens.After all at the bottom of the graph shows the confusion matrix in both form i.e. in minimum and in maximum form or we can also say that it also shows according to the cost and benefit.

The accuracy also shown in the **table.3** predicted by the table (a) and by table (b) classification accuracy in prediction (a) is

60.6666% and classification accuracy for prediction(b) is **30.3333%** shown in **table.4**and the result can also be different according to the population, target value, score threshold.

As, with the help of this algorithm we can see the graph and compare their results at a time. Both graphs can be opened and we can easily change their axis.

Axis	Attributes	Colour(marital status)
X	Education	Blue-single Red-married Black-divorced
Y	Income	

Table.1 Represents **Graph.1**

Axis	Attributes	Colour(education)
X	Marital status	Blue-high school Red-college yellow-graduate school
Y	income	

Table.2 Represents **Graph.2**

Threshold value	**0.75**
% of population	**10**
% of target	**25**
Classification accuracy	**60%**

Table.3 prediction of table (a)

Threshold value	**1**
% of population	**10**
% of target	**20**
Classification accuracy	**30%**

Table.4 prediction of table (b)

V. CONCLUSION AND FUTURE WORK

In this paper, the research work compares the result of all the algorithms with each other Naïve Bayes, J48 algorithm, ROC curve, pre-processing the data set. So it is found that the result of J48 tree and the ROC curve is far better or easy to understand compare to Naïve Bayes rule. This study is very helpful to identify the ratio of given or the collected data it can easily calculate the maximum, minimum, mean and standard deviation of the data. In future the result of Naïve Bayes multi functional can be improved because while performing the Naïve Bayes it does not make any effective difference by which result can be differentiate. So, in future there is scope that many other data mining algorithms can be added to it.

Method	Model Representation	Model Evaluation
J48 classifier	Tree,Matrixform	Very compact, easy to analyze
Naïve Bayes	Confusion Matrix	Posterior probability
Pre-process Data	Explicit & indicate all the data	Visualize all the data in bar form
ROC curve	Construction phase,curve,Graph	Classify accuracy, error, threshold values.

REFERENCES

[1]. P. Nevlud, M. Bures, L. Kapicak and J. Zdralek, "Anomalybased Network Intrusion Detection Methods

[2]. Advances in Electrical and Electronic Engineering, pp. 468-474, (2013.)

[3]. M. Phillips, "tcp2d," (18 November 2013). [Online]. Available: https://github.com/mrrrgn/tcp2d.

[4]. Charles A. Fowler and Robert J. Hammell Converting PCAPs into Weka Mineable Data copyright 2014 IEEE

[5]. SNPD 2014, (June 30-July 2, 2014), Las Vegas, USA.

[6]. M. Mayilvaganan, D. Kalpanadevi "Comparison of Classification Techniques for predicting the performance

[7]. Of Students Academic Environment", 2014 India, coimbotre International Conference on Communication and Network Technologies (ICCNT)

[8]. Patricia Morreale, Steve Holtz, Allan Goncalves, "Data Mining and Analysis of Large Scale Time Series Network Data", 2013 27th International Conference on Advanced Information Networking and Applications Workshops.

[9]. Eleonora Brtka*, Vladimir Brtka*, Visnja Ognjenovic* and Ivana Berkovic*," The data visualization technique in e-learning system", IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (September 20-22, 2012),

[10]. Sabri Serkan Güllüoğlu," Segmenting Customers With Data Mining Techniques", ISBN: 978-1-4799-6376-8/15/©(2015) IEEE.

[11]. Manisha Girotra, Kanika Nagpal, Saloni Minocha, Neha Sharma," Comparative Survey on Association Rule Mining Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 84 – No (10, December 2013)

[12]. C. M. Velu, K. R. Kashwan," Visual Data Mining Techniques for Classification of Diabetic Patients", Maharashtra, INDIA, 2013 3rd IEEE International Advance Computing Conference (IACC)

[13]. Swasti Singhal, Monika Jena," A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, (May 2014)

[14]. D. Rajeswara rao, Vidyullata Pellakuri, SathishTallam, T. Ramya Harika" Performance Analysis of Classification Algorithms using healthcare dataset" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015.

[15]. Darshana Parikh, Priyanka Tirkha,” Data Mining & Data Stream Mining – Open Source Tools” International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue (10, Octomber 2013)

[16]. S.Ummugulthum Natchiar, Dr.S.Baulkani,” Customer Relationship Management Classification using Data Mining Techniques” International Conference on Science, Engineering and Management Research (*ICSEMR* 2014)

[17]. Sandra Gama, Daniel Gonc¸alves,” Visualizing Large Quantities of Educational Data mining Information in 2014 18th International Conference on Information Visualization.

[18]. Pavel Blazek, Jiri Krenek, Daniel Jun, Krejcar, Ondrej, “The Biomedical Data Collecting System” Radioelektronika (RADIOELEKTRONIKA, 2015) 25th International Conference.