**Analysis of Classification Learning Algorithms**

Hana Rasheid Esmaeel

Department of information & communication, Al Nahrain Universit,Iraq

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Article history:*** |  | The paper attempts to apply data mining Technique, Five classification algorithms were used to build data they are (ZeroR, SMO, Naive Bayesian, J48 and Random Forest).The analysis implemented using WEKA (3.8.2) Data mining software tool. The information was collected from college of Information Engineering (COIE) In Al Nahrain University within the variety of form using "Referendum" to estimate the teacher performance; it was store in Excel file CSV format then regenerate to ARFF (Attribute Relation File Format). Many criteria like (Time taken to create models, accuracy and average error) was taken to evaluate the algorithms Random forest and , SMO Predicts higher than alternative algorithms ,since their accuracy is the highest and have lowest average error compared to others  ,"The teacher clarification and wanting to be useful to students " was the strongest attribute. Further removing the bad ranked attributes (10, 11, 12, and 14) that have a lower contact on dataset can increase accuracies of algorithms |
| ***Keywords:***  ***Data mining, Weka, Decision tree, classification, teacher evaluation*** |
| *Copyright © 2019 Institute of Advanced Engineering and Science.  All rights reserved.* |
|  | | |

1. **INTRODUCTION**

"Extracting the knowledge from the massive set of informationis calledData mining".[1][12],The

data ought to be new not obvious and one should be ready to use it ,that ought to be helpful when

deciding to find the buried patterns and interaction  that should be useful in deciding.

Data mining consists of **(5)** elements [1]

1. Extract, transform and load dealing Information.

2. Store and manage the info in an exceedingly two dimensional info systems.

3. Provide information access to business analysts and knowledge technology professionals.

4. Analyze the info by application code.

5. Reward the info in an exceedingly helpful format like a graph or table

Several works have used data mining to enhance the teacher performance; Asanbe M.O., Osofisan A.O. and

William W.F [2.] have "designed Artificial Neural Network (ANN) and Decision Tree scheme Their system was tested using data from a "Nigerian University", Renuka Agrawal1, Jyoti Singh, and Zadgoankar[3] "suggest a form to estimate the performance using data mining" like "association", classification rules ("Decision Tree", "Rule Induction"," K-NN", and "Naïve Bayesian") to find out traditions so as to assist them to enhanced supply the educational process "Hemaid and El-Halees[4] in their study they use questionnaire which contains a questions on the classes ,they intend a form to "evaluate teacher performance during the apple of data mining techniques like, classification, association rules to find out behavior so as to assist them toward enhanced the learning procedure and improve the presentation of teachers in classroom", to enhance the educational process and expand the contribution of teachers in the classroom". Works published by Ahmeda, Rizanerc and Ulusoyc [5] using Sequential Minimal Optimization ,Naïve Bayes , J48 Decision Tree ,and Multilayer Perception to Evaluate Student records to predicting the teacher performance and investigates factors that have affected students' achievements to develop the teaching system", [10] in his study to Predicted students' performance he finds that the classification scheme is repeatedly used in educational data mining area ,it include , Neural Network and Decision Tree ,the two methods greatly used by the researchers for predicting students' performance". Ms.A.Pavithra, Mr.S.Dhanaraj [11] In their study they examining the prediction accurateness of the academic performance of teaching the students using different classification algorithms like, "MLP, Naïve Bayes, Decision tree, REP tree, and J48 tree". They concluded that "many factors will influence the student performance and it may differ to the different locality of students".

1. **RESEARCH METHOD**

**2.1**. **Problem Definition:**

The problem is to evaluate the teacher performance in (COIE) In Al Nahrain University within the

Referendum to the students to estimate the teacher performance by using five classifications

Algorithms, then discover which attribute have the strongest effect in teacher evaluation, and

Which algorithm was the best?

**2.2.** Prepare data: For this analysis data was collected from college students at totally different

Departments within the faculty of College of Information Engineering (COIE) at Al Nahrain

University in Baghdad/Iraq, for the aim of investigation, however their skilled improvement

Has taken Place throughout the term. The info was collected from the college within the type

Of form to judge teacher performance as shown in Table (1) below. [3][4][5].

2.2.1. Preparing information for this analysis, teacher's data was evaluated. Attributes and value

Was determined.

2.2.2. Data saved in excel file in CSV (Comma Separated Values).

2.2.3. To use Weka data must be converted to Arrf (Attribute Relation File Format)

2.2.4. Using Weka (3.8.2) GUI Chooser and Explorer

2.2.5. Using Weka preprocessor and open Arff file.

2.2.6. Apply a Classification Algorithms (ZeroR, SMO, Naïve Bayesian, J48 tree and Random

Forest).

2.1.7. Evaluate the result and performance. Figure (1) below show the proposed system.

|  |
| --- |
| Data Collected From (COIE) In Al Nahrain University |
| Attribute And Value Are Determined |
| Data Saved In CSV File Format (Excel File) |
| Converted To Arff (Attribute Relation File Format) |
| WEKA ((3.8.2) GUI Chooser |
| Using Preprocessor |
| Apply Classification Algorithms (ZeroR, SMO, Naïve Bayesian ,J48 tree and Random Forest) |
| Evaluate Result And Performance |

Figure (1) the proposed system

Table (1) show the Questions and Their Abbreviations

|  |  |  |
| --- | --- | --- |
| **Seq** | **Attribute** | **Description** |
| 1 | DESG. | |  | | --- | | TITLE | |
| 2 | QUA. | Degree |
| 3 | EXP. | Experience |
| 4 | SC\_CTM\_ES | The semester course content pedagogic and analysis were provided at the beginning |
| 5 | CA\_OBJ.S | At the beginning The teacher were clearly specific the aims and purpose of the course. |
| 6 | CW\_A\_CA | The amount of credit allotted to  the course was positively  significance |
| 7 | CTA\_SA\_AY | The course was instructed in step with the information proclaimed on the primary day of sophistication |
| 8 | CD\_HW\_ASS.\_APP.\_SAT | The class discussions homework assignment applications and studies were satisfactory |
| 9 | TB\_other\_CR | The text book and different courses resources were enough and up to this point |
| 10 | C\_WAPP.\_LAB\_DIS | The course acceptable conversation of the laboratory applications and different studies. |
| 11 | QUIZ.ASS.PROJ.EXA.\_HEP. | The quizzes assignment comes and exams contributed to serving to the educational. |
| 12 | ENJ.CLA.ACTIVI.DUR.LEC. | The lecture allows students to participate their knowledge. |
| 13 | INT.EXPE.C.END.Y | The course were met all student prospect |
| 14 | CWR\_BTM\_PRO.\_DEVE | The course was relevant and helpful to my skilled development. |
| 15 | CHM\_LA\_W\_PRE. | The course helped Maine check up on life and world with my new perspective. |
| 16 | INS.\_KW\_RELE.\_DATE | The lecturer information was relevant and up to this point. |
| 17 | INS.\_CP\_clas | The lecturer came ready for categories |
| 18 | INS.\_TAUG.IN\_ACCOR.\_ | The lecturer instructed in accordance with the proclaimed lesson set up. |
| 19 | INS.\_W\_COMMI.\_T\_THE.\_CO. | The lecturer was committed to the course and was comprehendible |
| 20 | INS.\_ARR.\_OF\_T.C | The lecturer arrived of your time for categories |
| 21 | INS.\_H.SM.\_C\_HO. | The lecturer had a sleek associate of sophistication hours |
| 22 | INS.\_EXP.\_THE\_CO.\_A.W.E.\_HE.\_T\_ | The teacher clarification and was wanting to be useful to students |
| 23 | INS.\_DE\_AP\_APP.\_T\_ST. | The lecturer incontestable appositive approach to students |
| 24 | INS.\_W.OP.\_RES.\_VIE.\_ST.\_CO. | The teacher was considerate of the views of student on the topic of the course. |
| 25 | INS.\_EN.\_PART.\_IN\_CO. | The lecturer inspired participation within the course. |

**2.3.** The knowledge was gathering for making ready the model, the fields that are needed for data processing was taken, This includes pre-processing or extracts vital info from it then produce correct format file of the info like in weka.arff file format (Attribute Relation File Format) [3] ,as shown in Fig **(2)** below.

@relation *teacher evaluation*

@attribute *Name1* string

@attribute *Name2* numeric

@attribute Namen?

@data

*Yes, 1, yes….*

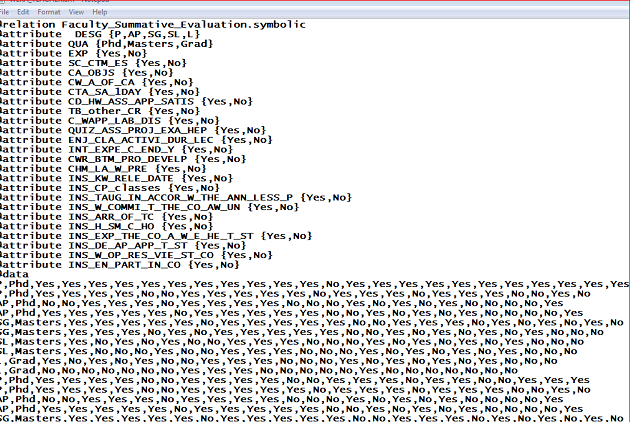


Figure (2) below show( teacher.arff) file

**2.4.Methodology**

The classification technique was used for the forecast of teachers’ evaluation... Five classification algorithms was used (ZeroR, SMO, Naïve Bayesian, J48 tree and Random Forest) and implemented using WEKA (3.8.2) Data mining software tool.

**2.5.Building Models**

**2.5.1.Building the Trivial model ZeroR [7] [8] [10]**

**2.5.1.1**.In Preprocess” panel, click “Open file” button, and choose file named (teacher.arff) as

in figure(3).

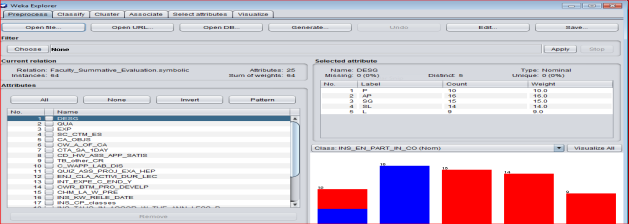


Fig (3) selected attributes of teacher.arff

**2.5.1.2.** Select ZeroR by Clicking “Choose” Button.  
 **2.5.1.3.** Invoke classifier by clicking “start” button to make a model. The analytical performance of

The model characterized within the right hand Classifier output frame. [5]   
 **2.5.1.4.** The Confusion Matrix for the model is bestowed at the underside a part of the Classifier output

Window. It is seen from it that compounds are classified as (21) affirmative and (43) No   
 **2.5.1.5** the accuracy of the model is (67.187) for No and (32.813) for affirmative as in figure (4) and

Figure(5).

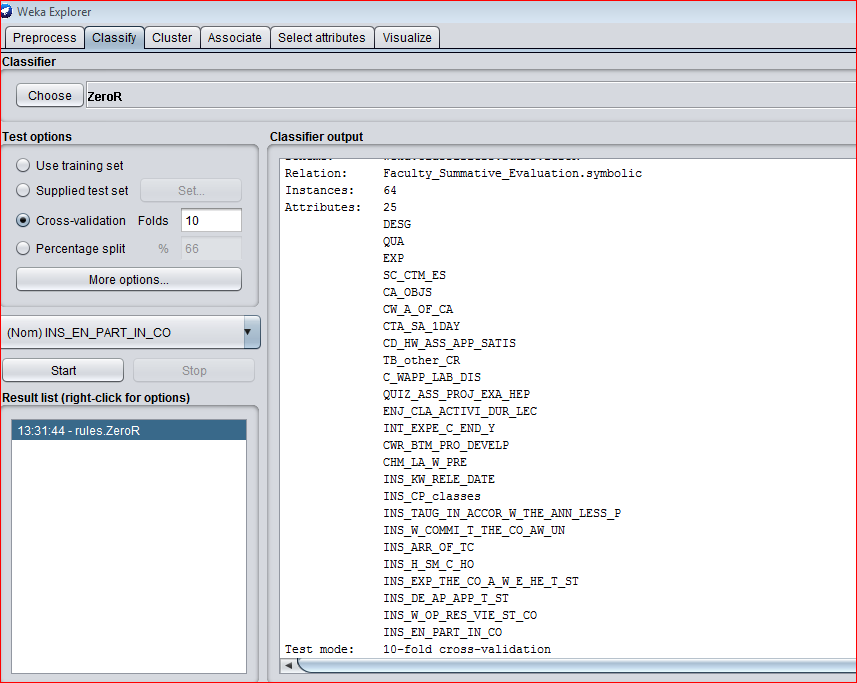


Fig. (4) ZeroR classifier output

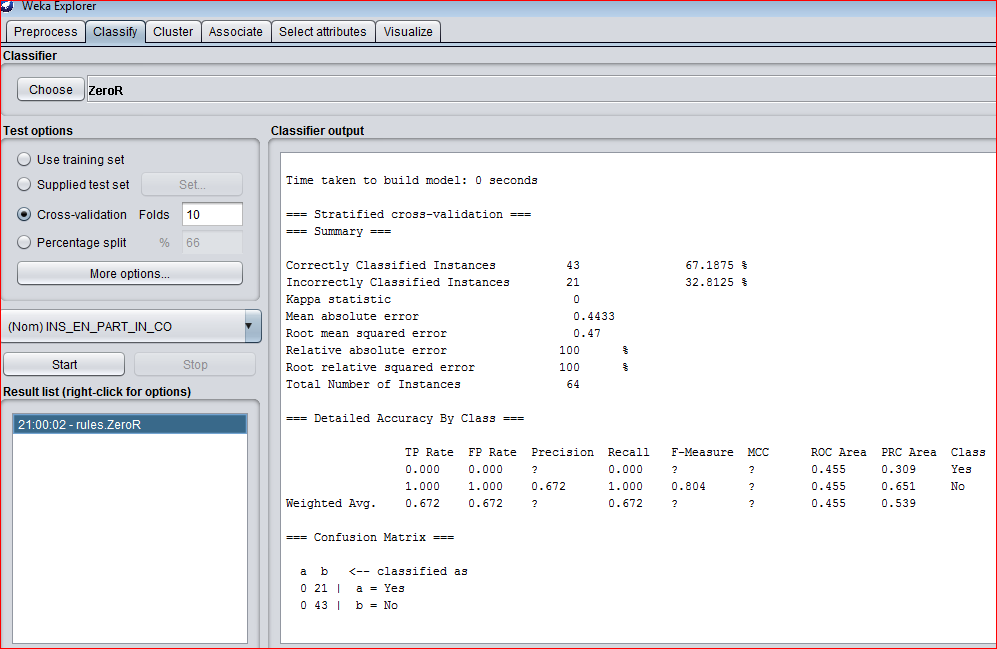


Fig (5) continue of ZeroR classifier output

**2.5.2.Building J48 Tree Model: [5] [6]**

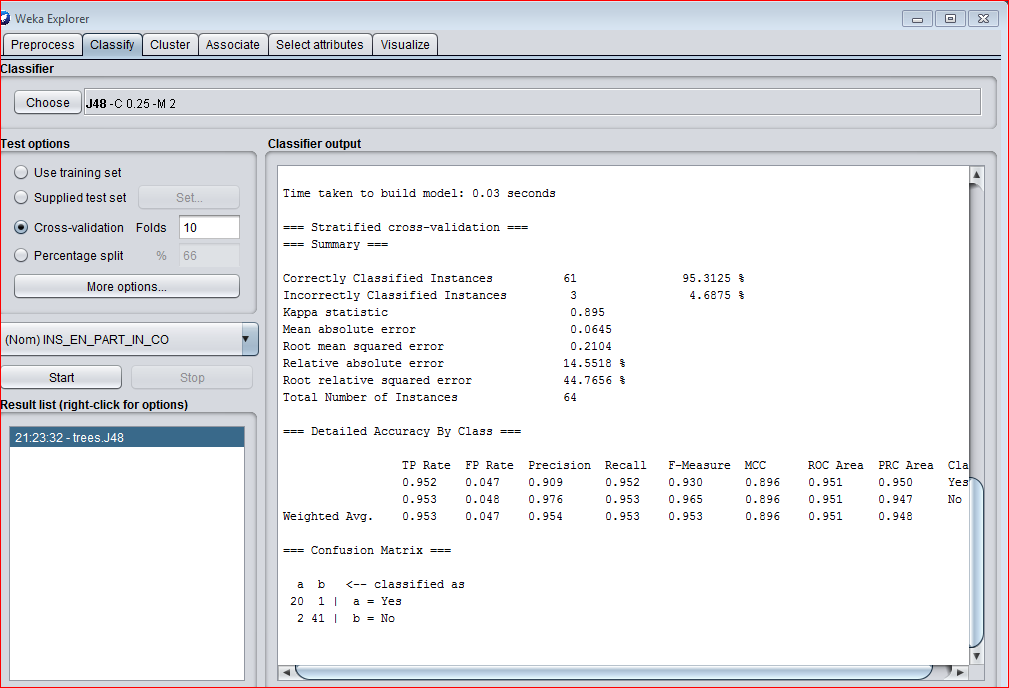
In order to get usual illustration of the tree, do the following:

2.5.2.1.Click the correct push on the model Type trees(J48) within the Result list frame and choose the

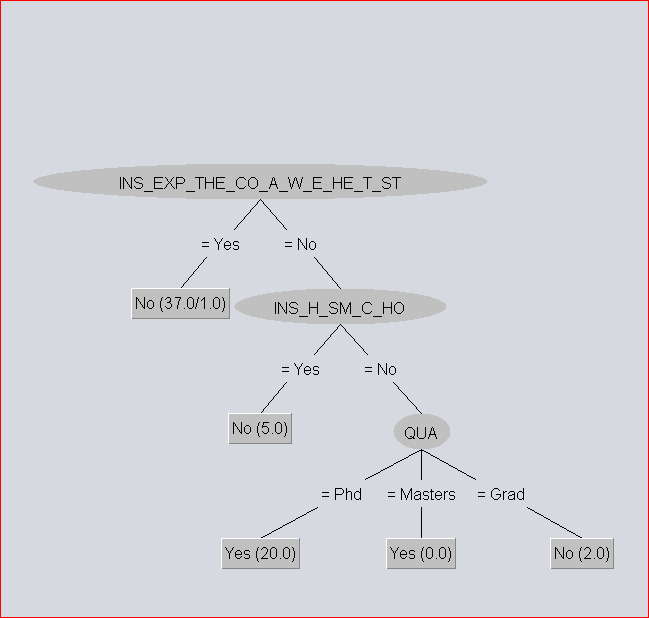
menu item Visualize tree Size a replacement window with graphical Illustration of the tree.

2.5.2.2.Click with the correct push to the Area during this screen, and within the popup menu Choose the

item appropriate screen. As in Figures (6) and (7) below.

****

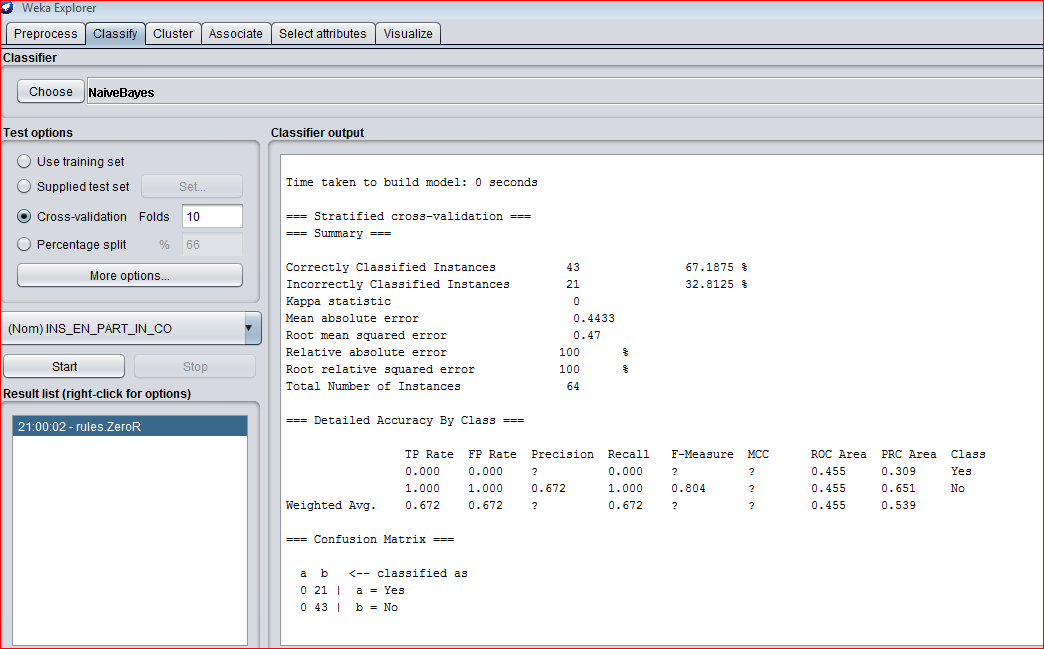
**Fig (6) the Statistical of the J48 model.**

****

**Fig (7) Visualize j48 tree**

**2.5.3. Building Naive Bayesian Model:** As in previous models Naive Bayesian Model was built

As shown in figure (8)

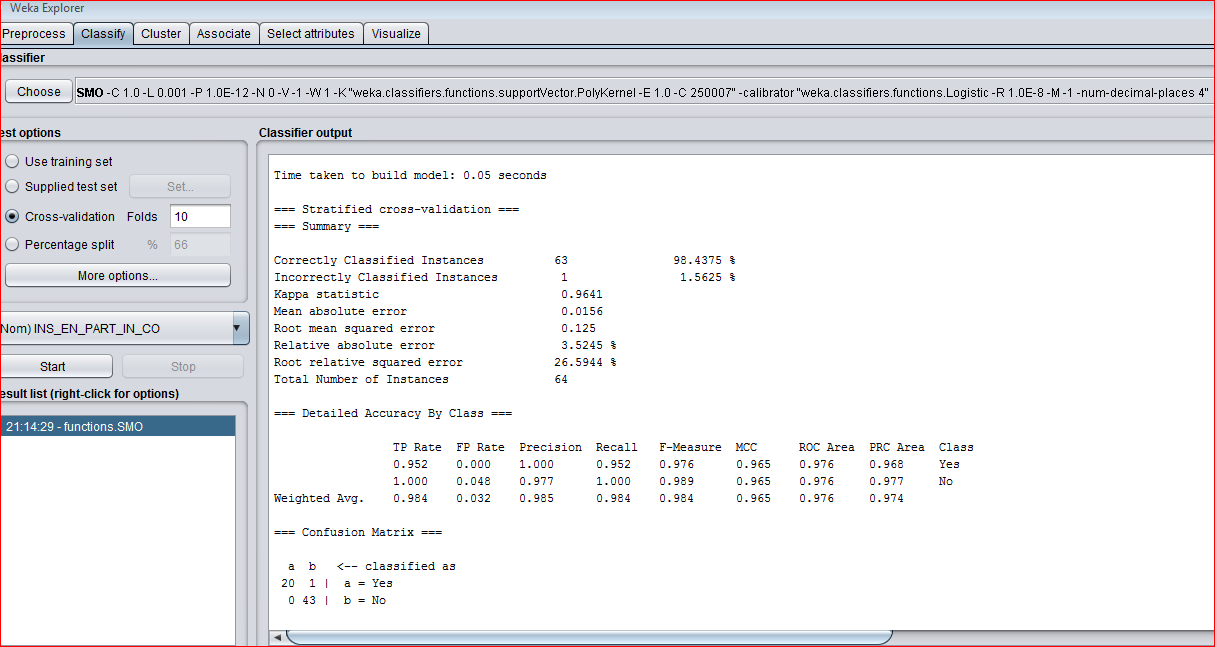
****

**Fig. (8)**  **Navie Bayesian Model**

**2.5.4. Building Support Vector Machine Models:[6] [7]**

The Weka software implements Sequential Minimal Optimization (SMO) algorithm for training a

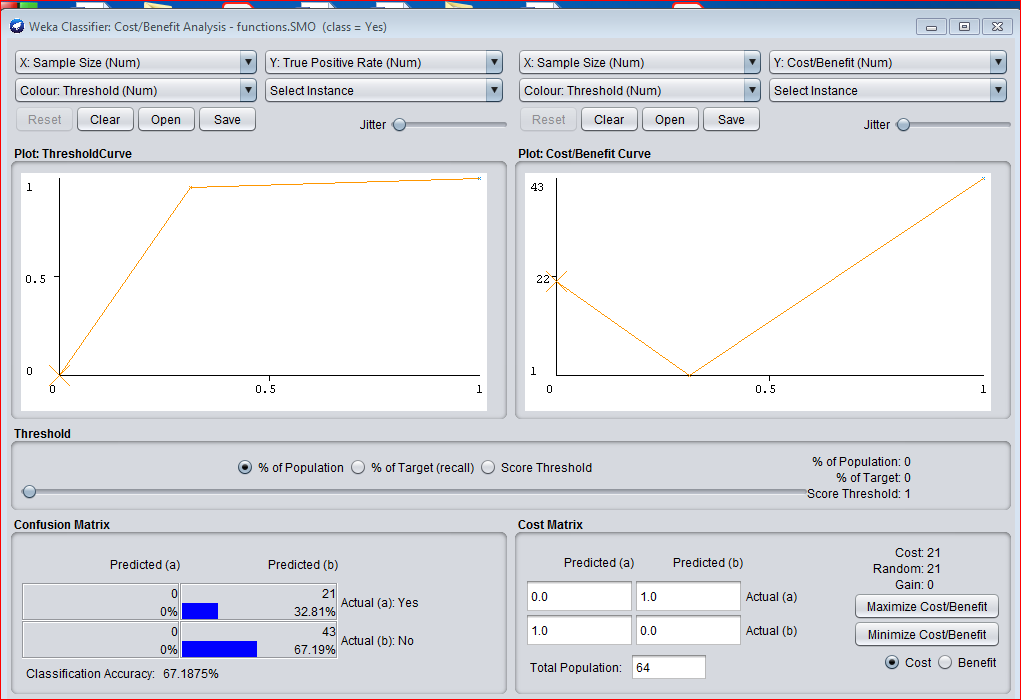
Support vector classifier, Figure (9) below show the Classifier output for this model.

****

**Fig (9) the classifier output for SMO**

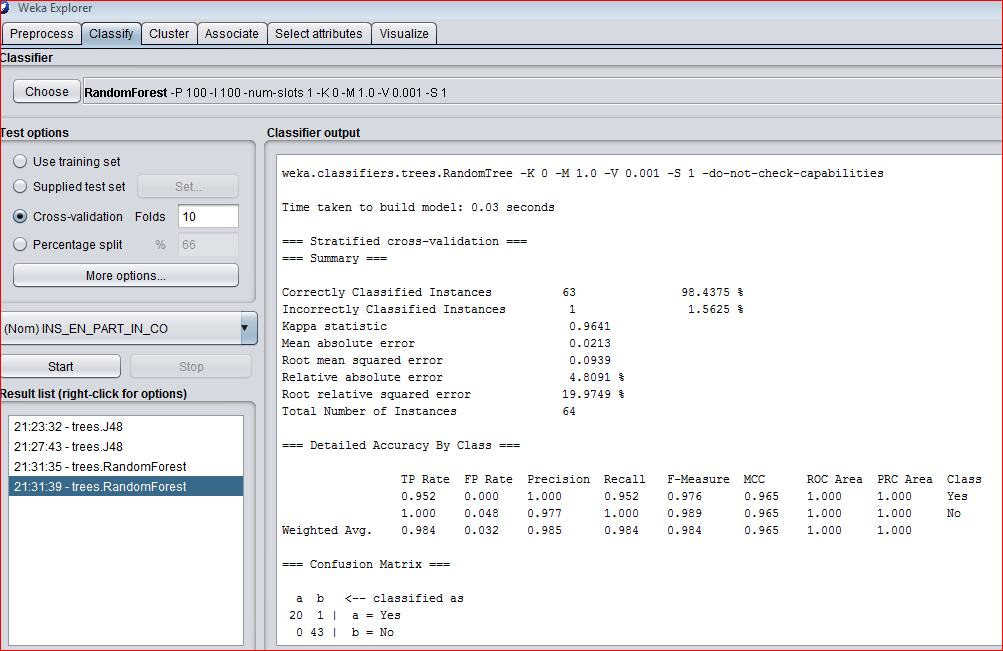
The accuracy of (Correctly Classifieds Instances) of this model is extremely high ninety eight.4375%. This truth clearly indicates that the accuracy can't be used for assessing the utility of classification models designed exploitation unbalanced datasets. For this purpose an honest selection is to use the “Kappa statistic”, that is =0.964 for this case. “Kappa statistic” is associate degree analog of the coefficient of correlation. It's worth is zero for the shortage of any relation and approaches to (1) for terribly sturdy applied

math relation between the category label and attributes of instances, Another helpful applied math characteristic is “ROC Area”, that the worth =0.976 means that sensible mythical monster curves may be build and therefore the cost/benefit analysis will simply be performed. As in figure (10).

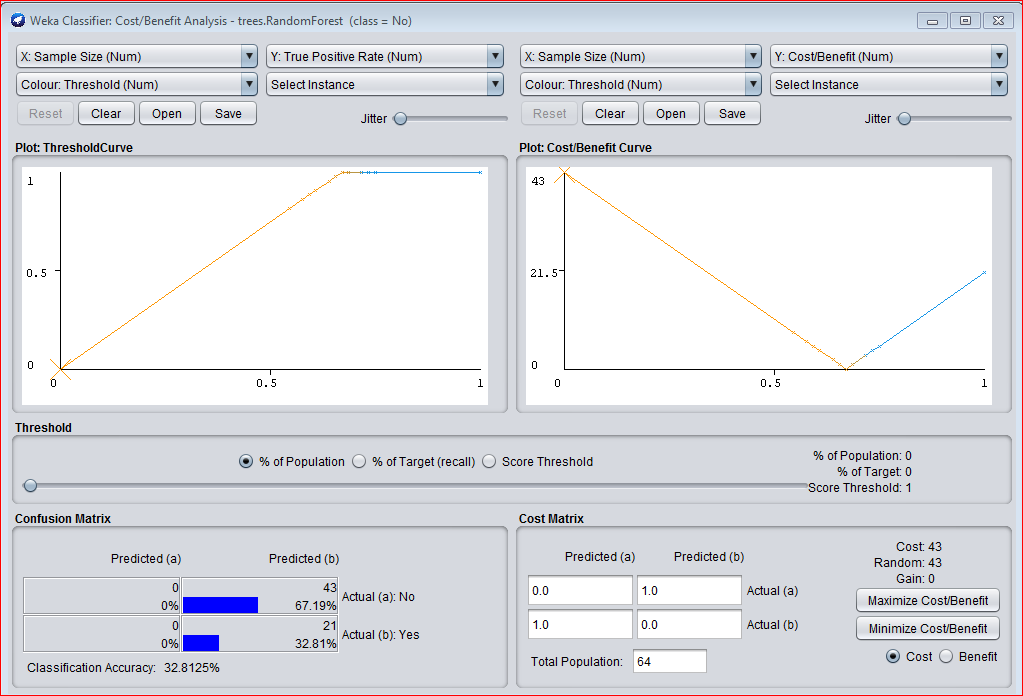


**Fig. (10) cost/benefit analysis for SMO**

**2.5.5**. **Building the Random Forest Models** [8] [10]:The classifier output used for "Random Forest" algorithm be shown in Figure (11) below the accuracy of this model is extremely high= ninety eight.4375%. This reality clearly indicates the accuracy cannot be used for assess the worth of classification models designed victimization unbalanced datasets. For this purpose an honest alternative is to use the “Kappa statistic”, that is =0.964 for this case [6]. Its price is =0.9641 it's terribly robust applied math relation between the category label and attributes of instances, Another helpful applied math characteristic is “ROC Area”, that the worth =1.000 means that sensible mythical creature curves are built and therefore the cost/benefit analysis will simply be performed. As in figure (12) .

.****

**Fig. (11) Classifier output for Random Forest**



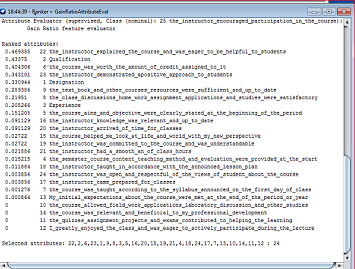
**Fig. (12) cost/benefit Random Forest**

**2.5.6.** **Attribute Ranking**: Weka explorer can evaluate the attributes of the data by using the following:

Select attributes→GainRatoAttribute→start→show Eval→Rank Attribute

The list of attributes and their value are appear from the higher to lower.as in table (2) below.

Table 2: Attribute Ranking



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2.6. Models Comparison:**  The performances of the (**5**) models were evaluated primarily based on the standards as illustrated  in table (3) below.   1-Prediction accuracy :The share of  properly classified instances is usually referred to as accuracy  of a model.   2- Time taken to create the model.  3- Error rate.  **Table3 :Comparison analysis on the models**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Random  Forest | SMO | Naive  Bayesian | J48 | ZeroR | Metric | | 0.03 | 0.05 | 0 | 0.03 | 0 | Time To Build The Model | | 98.435% | 98.437% | 92.187% | 95.312% | 67.187% | Correctly Classified Instances | | 1.562% | 1.562% | 7.812% | 4.687% | 32.812% | In Correctly Classified Instances | | 0.964 | 0.964 | 0.833 | 0.895 | 0 | Kappa Statistics | | 0.0213 | 0.0156 | 0.0791 | 0.0645 | 0.443 | Mean Absolute Error | | 0.0939 | 0.125 | 0.279 | 0.210 | 0.47 | Root Mean Square Error | | 4.809 % | 3.524% | 17.838% | 14.55 % | 100% | Relative Absolute Error | |  | 26.594% | 59.428% | 44.765% | 100% | Root Relative Square Error | |  | PhD | Masters | Grad |  |  |

**2.7.** **Performance of the models**: Table(4) show the performance of the (5) algorithms Where**:[7][8]**

**TP**=true positives": variety of examples":Predicted positive that are literally positive.  
 **FP**=false positives" : variety of examples:"Expected positive that are literally negative.  
 **TN**=true negatives": variety of Examples ":predicted negative that are literally negative.  
 **FN**=false negatives" : variety of Examples :"Expected negative that are literally positive.  
 **Weka (3.8.2) Confusion Matrix**: The quantity of  properly classified instances is that the total of   
 diagonals within the matrix; all Others area unit Incorrectly classified.  
 x y <-- classified as , actual x=0 TP FN  
 actual y=1 FN TP  
 TP=TP+FN / Recall   
 Precision=TP/TP+FP  
 Accuracy=TP+TN /TP+TN+FP+FN

**Table 4: performance of the models**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Roc Area | F-Measure | Recall | Precision | FP  Rate | TP Rate | Algorithm |
| 0.455 | ? | 0.672 | ? | 0.672 | 0.672 | ZeroR |
| 0.951 | 0.953 | 0.953 | 0.954 | 0.047 | 0.953 | J48 Tree |
| 0.927 | 0.924 | 0.922 | 0.937 | 0.038 | 0.922 | Naive Bayesian |
| 0.976 | 0.984 | 0.984 | 0.985 | 0.032 | 0.984 | SMO |
| 1.000 | 0.984 | 0.984 | 0.985 | 0.032 | 0.984 | Random  Forest |

1. **CONCLUSION :**From the result of Comparison of the five algorithms as in tables (4) and (5) it conclude that Algorithms SMO and Random forest Predicts higher than alternative algorithms since theier Accuracy is that the highest and have lowest average error compared to others algorithms . On functioning on performance, several attributes are tested, and found that a few of them are Effective on the performance Prediction. "The teacher clarification and was wanting to Be useful to students" was the strongest attribute and then the result plays A vital role within the performance of academics .More a lot of removing the worst hierarchal attributes (10, 11, 12, And 14), that have a lower impact on dataset can increase the Algorithms performance accuracies.

**REFERENCES**

[1]Shafiq Aslam,Imran Ashraf," Data Mining Algorithms and their applications in Education Data Mining", International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 7, July 2014 pg. 50-56

[2] Asanbe M.O., Osofisan A.O., William W.F. "Teachers’ Performance Evaluation in Higher Educational Institution using Data Mining Technique", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 10 – No.7, March 2016 – [www.ijais.org](http://www.ijais.org)

[3] Renuka Agrawal1, Jyoti Singh, A.S. Zadgoankar, "Summative Assessment for Performance Evaluation of a Faculty Using Data Mining Techniques ", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 10, October 2016 Copyright.

[4]Randa Kh. Hemaid1, Alaa M. El-Halees,"Improving Teacher Performance using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015.

[5]Ahmed Mohamed Ahmeda, Ahmet Rizanerc, Ali Hakan Ulusoyc,"Using data mining to predict instructor performance", 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS , 29-30 August 2016, Vienna, Austria.

[6]Wikipedia.org<http://en.wikipedia.org/wiki/Weka_(machine_learning)>

[7]Tutorialspoint.com, Data Mining Tutorial Simply Easy Learning, 2-11-2014

[8]<http://weka.wikispaces.com/Primer>

[9]The WEKA Workbench Eibe Frank, Mark A. Hall, and Ian H. Witten Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, Fourth Edition, 2016

[10] A.Mohamed Shahiria,, W. Husaina , N. Abdul Rashida, "A Review on Predicting Student’s Performance using Data Mining Techniques" Procedia Computer Science 72 ,414 – 422, ELSEVIER 2015. Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

[11]Ms.A.Pavithra, Mr.S.Dhanaraj,"Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors", IJSRCSAMS Volume 7, Issue 5 (September 2018) [www.ijsrcsams.com](http://www.ijsrcsams.com).

[12]Miss. Sharayu N. Bonde , Dr. D. K. Kirange, "Survey on Evaluation of Student's Performance in Educational Data Mining", Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974