IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Analysis of PG Admission in Arts and Science College using Data Mining Tools

P. Sundari

Department of Computer Science, National College, Tiruchirappalli, Tamil Nadu, India

*Corresponding Author: periasamy_sundari@yahoo.co.in

Available online at: www.ijcseonline.org

***Abstract*—** Information is the backbone of any Organization. To succeed in this situation, one must manage the information in the right amount and at right time. Data Mining is used to mine the new pattern or trends or rules from the unknown/ large amount of data / unpredictable data sets. Data Mining is used at diverse fields like Educational, Agriculture, Medical, Police Department, Research side, Information Technology side and Image processing etc., This Paper analyzes the mindset of Final year UG student about PG admission. For that 18 attributes and one class label collected from the final year UG students then applied on the Data Mining Tools like Weka Tool and Orange Tool. The data sets passed on the classification algorithms like J48, Naive Bayes, RandomForest and REPT Tree in Weka and Classification Tree, CN2, Naive Bayes and kNN of Orange Data Mining Tool. The Confusion matrix, Training and simulated Errors and Testing and Validation Results are obtained and tabulated. The Weka and Orange data mining tools classifiers performance are represented in the graphical form and its decision tree. From the decision tree, hidden rules are extracted, from which possible to determine factors which affects the PG admission. From the Weka tool, obtain attributes interest, jobavailability, feestat, colinfra and scholarship can predict the PG admission. In orange data mining tool, attributes interest to study, jobavailability, feestatus, scholarship, gender, pregovexam and colinfra from the original data set can predict the PG admission.

***Keywords*—** *Data Mining, Weka, Orange Data Mining Tool, Classification Algorithms, Evaluation Result, Confusion Matrix, Rules.*

I. INTRODUCTION

Data Mining can be applied on diverse areas like Business and E-Commerce, Scientific, Engineering and Health Care, Web data etc. Utilizing the techniques, a growing body of applications is emerging i.e. changing the landscape of business decision support. To mine or extract information from various large data sets as industrial or business data set various data analysis algorithms have been developed. These data mining algorithms have been categorized into three types: Predictive Modeling, Clustering and Frequent Pattern Extraction. In this paper, classification algorithms of data mining tools are taken for consideration. A decision tree is a flow-chart like tree structure. For that chosen data mining tools are Weka and Orange data mining tool. Both are open source software.

II. RELATED WORK

To do the research same analysis of admission process carried out by many researchers, most of the notable research works mentioned below:

Rakesh Kumar Arora, Department of Computer Science, Krishna Engineering College, Ghaziabad, UP, India, Dr. Dharmendra Badal, Dept. of Mathematical Science & Computer Applications, Bundelkhand University, Jhansi, U.P, India, "Admission Management through Data Mining using WEKA"[1], a simple methodology based on k-means clustering algorithm is being used to analyze the data obtained from the admission form filled by admission seekers. This methodology will assist the academic planners to monitor admission details of students seeking admission in institute over the years. Hence this model will play important role in determining the reasons for decline in quality of admissions taken in the institute over the year and steps that need to be taken to improve performance from next academic session. This model will help in identifying the set of students that need to be focused to actually convert the inquiry into admission.

"Mining higher educational students data to analyze student's admission in various discipline" – they collected data from Alpha College and applied on data mining techniques and result calculated. For that they collected details about Course, Branch, Gender, Class, Income, Date of Admission and Minority / Non-Minority. Decision Tree, Cluster Analysis and Association rule data mining technique used on the Weka Tool. The data mining techniques that

used are association rules and decision tree help to improve the result. They help to uncover the hidden patterns from the large data. These hidden patterns are useful for both teachers and management. Decision trees are generated depending various attributes such as course, gender, exam board name, class obtained by students, date of admission, category, minority/nonminority.

III. METHODOLOGY

This paper focus on applying the student data set into the classification algorithms on the data mining tools Weka and Orange. Then extract the hidden rules which helpful for improving the PG admission. The methodology of study consists of three steps:

a) the selection of Data Mining Tools
b) Data sets to be used
c) the selection of Classification Algorithms.

The selection of Data Mining Tools: For this paper, Weka and Orange data mining tools selected. Weka [Waikato Environment for Knowledge Analysis] is open source software under the GNU General Public License. System is written using Object Oriented Language Java and developed at the University of Waikato in New Zealand. Weka has a collection of machine learning algorithms for data mining tasks and also contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Orange data mining tool is a component based visual programming software for data mining, machine learning and data analysis. This tool is written in Python, Cython, C++ and C. No need for programming. Easy to Learn. Data Mining is done through Visual Programming or Python Scripting. Users can run their Python scripts in a terminal window, integrated environments like PyCharm and PythonWin or Shells like iPython. Orange contains canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing the data table, selecting features, training predictors, visualizing data elements and comparing learning algorithms. User can interactively explore visualizations or feed the selected data subset into other widgets.

Data set Used: The data was collected from final UG students in Arts and Science College. The numbers of students involved in analysis are 160 and parameters analyzed in the paper include: gender,fedu,medu, parentst, fsupport, relbwfam, freetime, wstat, hstat, feestat, parentgovemp, jobavailability, colinfra, pregovexam and class label joinpg. For Weka Tool – all records stored in .arff(attribute relation file formant) format. For Orange Tool – all records entered in Excel with .tab file format. Data preprocessing needed for data analysis which helps to remove the noise from the dataset. In weka tool, preprocessing tool used for removing noise. In orange tool, need to use projection Linear projection and Outlier.

The selection of Classification Algorithms: In Weka, classification algorithms: J48, Random Tree, REPTree and Naïve Bayes Algorithm applied on the collected data set. In Orange, classification algorithms Classification Tree, CN2, Naïve, Random Forest and K-Nearest Neighbor applied on the collected data set.

IV. RESULT AND DISCUSSION

The possible values of parameters considered in this paper are shown in the following

Table 4.1

Parameters	Description	Possible Values
gender	candidate's gender type	male, female
mstat	marital status	yes,no
fedu	father educated	yes, no
medu	mother educated	yes, no
parentst	parent status	living together, separated
fsupport	whether family support for education	yes,no
relbwfam	relationship between family members	excellent, good,notgood
freetime	after college hour he/she have a free time	yes,no
wstat	after he/she working?	yes,no
hstat	student health status	ok,notok
feestat	fees status	ok,notok
parentgovemp	parent government employee?	yes,no
scholar	got scholarship?	yes,no
jobavailability	job availability exist?	yes,no
interest	really having interest to study?	yes,no
colinfra	about college infrastructure	ok,notok
pregovexam	preparing for government job?	yes,no
alumni	alumni suggestion	yes,no
joinpg	join pg course	yes,no

In Weka Tool

The number of instances used for training was 76 and used attributes 19. In weka, used classification algorithms are J48, Random Forest, REPT Tree and Naïve Bayes.

Table 4.2 Training and Simulation Error Table: [In Weka]

Algorithm	Kappa Statistics	Mean Absolute Error	Root Mean Squared error	Relative Absolute Error	Root Relative square error
J48	0.77	0.1399	0.2645	29.6%	54.44%
Random Forest	0.473	0.25	0.5	52.845%	102.85%
REPT Tree	0.440	0.306	0.4419	64.69%	90.915%
Naïve Bayes	0.661	0.1855	0.3023	39.25%	62.22%

The J48 decision tree's value of RMSE and RRSE has least value when compared to the values of the other models. This result reveals that decision tree is more suitable for the prediction of student's pg admission.

Testing and Validation Results:

The models obtained from the training data were rerun using the test and validation data sets to evaluate the performance of the resultant models. A total of 76 instances were used.

Table-4.3- The following Table-4.3 shows the evaluation measures (confusion matrix):

Classifier	**Correctly Classified Instances**	**Incorrectly classified instances**
J48	68	8
Random Forest Tree	57	19
REPT Tree	55	21
Naïve Bayes	64	12

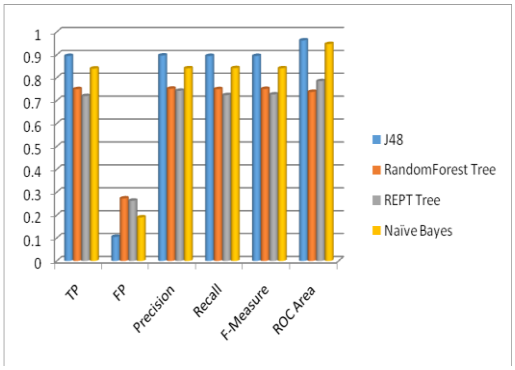
The table shows that the number of correctly classified instances of J48 decision tree is higher compared to other models. So the decision tree model plays an important role in predicting the pg course admission.

The study also evaluated the performance of the classifiers using the following metrics: Classification Accuracy, Miscalculation Rate, ROC Area, Precision and Speed. The performance measures of the classifiers using 10-fold Cross Validation results are tabulated as below Table4.4:

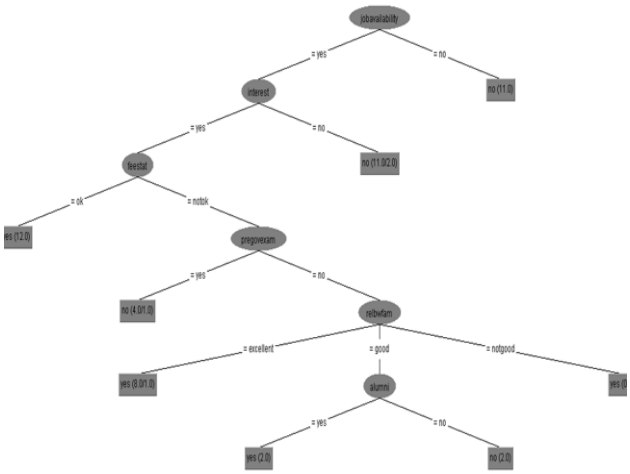
Table4.4:

Algorithm	TP	FP	Precision	Recall	F-Measure	ROC Area
J48	0.895	0.105	0.897	0.895	0.895	0.963
Random Forest Tree	0.75	0.273	0.752	0.75	0.751	0.738
REPT Tree	0.72	0.263	0.743	0.724	0.727	0.785
Naïve Bayes	0.84	0.19	0.841	0.842	0.841	0.947

Graphical representation of Weka tool's various classifier performance:



J48 Decision tree:



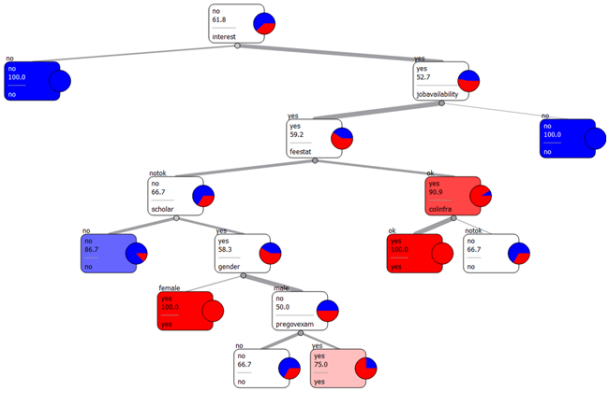
:

Extracted Rules:

interest = yes
| jobavailability = yes
| | feestat = ok
| | | colinfra = ok: yes (19.0)
| | | colinfra = notok: no (3.0/1.0)
| | feestat = notok
| | | scholar = yes: yes (12.0/5.0)
| | | scholar = no: no (15.0/2.0)
| jobavailability = no: no (6.0)
interest = no: no (21.0)

In Orange Tool

The following classification tree generated using orange tool:



Extracted Rule:

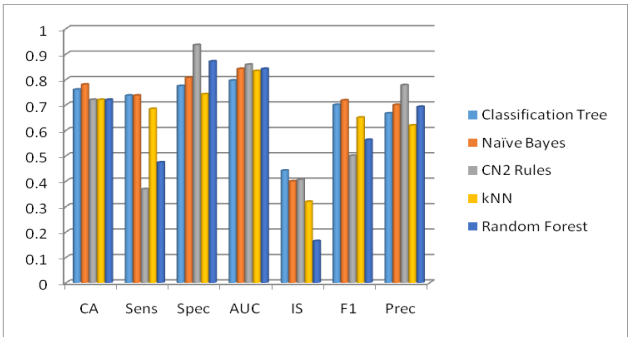
1. If interest = no then joinpg=no
2. If interest = yes and jobavailability = no then joinpg=no
3. If jobavailability=yes and feestatus=ok and colinfra = ok then joinpg=yes
4. If jobavailability = yes and feestatus = ok and colinfra=not ok then joinpg=no
5. If jobavailability = yes and feestatus=notok and scholar=no then joinpg=no
6. If jobavailability = yes and feestatus=notok and scholar=yes and gender=female then joinpg=yes
7. If jobavailability = yes and feestatus=notok and scholar=yes and gender=male and pregovexam=yes then joinpg=yes
8. If jobavailability = yes and feestatus=notok and scholar=yes and gender=male and pregovexam=no then joinpg=no.

The number of instances used for training was 76 and used attributes 19. In Orange tool, used classification algorithms are Classification Tree, Naïve Bayes, CN2 Rules, kNN and Random Forest. Its evaluated results are specified in the following Table 4.5.

Table 4.5

Algorithm	CA	Sens	Spec	AUC	IS	F1	Prec
Classification Tree	**0.7600**	**0.7368**	**0.7742**	**0.7958**	**0.4416**	**0.7000**	**0.6667**
Naïve Bayes	**0.7800**	**0.7368**	**0.8065**	**0.8417**	**0.3994**	**0.7179**	**0.7000**
CN2 Rules	**0.7200**	**0.3684**	**0.9355**	**0.8583**	**0.4053**	**0.5000**	**0.7778**
kNN	**0.7200**	**0.6842**	**0.7419**	**0.8333**	**0.3190**	**0.6500**	**0.6190**
Random Forest	**0.7200**	**0.4737**	**0.8710**	**0.8417**	**0.1642**	**0.5625**	**0.6923**

From the above table, classification accuracy of Classification tree (0.7600) and Naïve Bayes (0.78) comparing with other models. Classification tree representation like a decision tree, rules also formed and it is easy to understand. So Classification Tree model is an important to predict the admission of PG course.



V. DISCUSSION OF THE RESULTS

The data set collected from the final year UG students applied on the weka tool which contains various classification models like J48, Naïve Bayes, Random Forest and REPT Tree. The precision of J48 classify algorithm is 0.897 which is a higher value compared with the rest of the classification model. The original data set contains 18 attributes and one class label which is applied on the J48 Tree. The attributes interest, jobavailability, feestat, colinfra and scholar predict the PG admission. In orange tool, classification accuracy of Classification Tree(0.7600) and Naïve Bayes (0.7800) is high compared to the rest of the models like CN2, Random Forest and kNN. From the Classification Tree, certain attributes interest, jobavailability, feestat, scholar, gender, pregovexam and colinfra from the original data set can predict the PG admission.

VI. CONCLUSION

This paper analyzes the PG admission, by collecting the data set from the final year UG students. Using the Weka tool, obtained result is if the student has interest and job availability is exist and fee status is ok and college infrastructure is ok and scholarship is exist then student will to join PG course. If job availability is no or interest is no or fee status is not ok or college infrastructure is not ok or scholarship is not exist then student not willing to join PG course. From the Orange data mining tool, predict the Admission of PG Course by student have interest to study and job availability is exist and fee status is ok and scholarship exist and student is female then willing to join PG course. If the student prepare for government exam then willing to join PG course. If the student study interest is no or job availability is no or fee status is not ok or scholarship is no or prepare for government exam is no or college infra structure is not ok then students status of joining PG course is no. By using the extracted rules, make certain changes which cause the increases in the PG Admission.

REFERENCES

[1] Rakesh Kumar Arora, Department of Computer Science, Krishna Engineering College, Ghaziabad, UP, India, Dr. Dharmendra Badal, Dept. of Mathematical Science & Computer Applications, Bundelkhand University, Jhansi, U.P, India, "*Admission Management through Data Mining using WEKA*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[2] Neeraj Bhargava, MDS University, Anil Rajput, Govt, PG Nodal College, Sehore (M.P) India, Pooja Shrivastava, Research Scholar Barkatullah University, Bhopal, "*Mining higher educational students data to analyze student's admission in various discipline*", Binary Journal of Data Mining & Networking 1 (2010) 01-05.

