

Comparative Analysis of Decision Tree Algorithms for the Student's Placement Prediction

Er.Paramjit kaur¹, Er. Kanwalpreet Singh Attwal²

M.Tech Scholar, Department of Computer Engineering, Punjabi University, Patiala, India ¹

Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, India ²

Abstract: This proposes paper apply decision tree algorithms to predict the result of student campus placement. In this paper we also perform the comparative analysis of three decision tree algorithms (Id3,J48and random tree).In our study important classification model is generated to decide which the students might placed or non placed during campus placement process. The experiment is conducted using WEKA open source data mining tool on the real data set collected from final year students of computer engineering department of Punjabi University ,Patiala.

Keywords: Decision tree,Id3,J48,Random tree, WEKA,KDD

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amount of data [1]. It is analogous to finding gold from the rocks or sand. Data mining is extraction of useful patterns from various data sources, e.g., databases, texts, web, image. It allows us to find the needles hidden in our haystacks of data. It refers to using a variety of techniques to identify information in the bodies of data. Extracted information can be put to use in the areas such as decision support ,prediction, forecasting and estimation.

“Educational data mining is an emerging discipline concern with developing methods for exploring the unique type of data that come from educational setting and using those methods to better understand students and the setting which they learn in” as defined by the educational data mining community [2] .

Education is an essential element for the progress of country. Mining in educational environment is called educational data mining. It is concerned with developing new methods to discover knowledge from educational database [3]. Educational data mining provides a set of techniques, which can help the educational system to overcome these issues [4].The various techniques of data mining

are clustering ,classification ,prediction ,association rule etc. Now a day's use of EDM is increase ,due to increase in educational data set.

This paper ,introduces the use of classification technique of data mining for Engineering student's campus placement. In this research a large number of attributes are used to determines whether the student will be placed or not. Classification is used find the association among these different attributes and their affects on the student's campus placement.

The main purpose of the study to identify rules to predict the student 's placement and do the performance comparison of ID3,J48,Random tree.

In this research Random tree (decision) algorithm of classification is used to generate rules for student's

placement prediction. Data mining tool WEKA is used for rules generation .Then generated rules can be used for developing placement prediction application .

Decision trees are a simple, but powerful form of multiple variable analysis. A decision tree is a special form of tree structure. The tree consists of internal nodes where a logical decision has to be made, and connecting branches that are chosen according to the result of this decision.

The nodes and branches that are followed constitute a sequential path through a decision tree that reaches a leaf node (final decision) in the end.[10] .Classification has been used to provide new, important and therefore demand-oriented impulses for the development of new bachelor and master courses .

This technique is useful to identify Students' failure patterns , to determine parameters related to the admission process, migration, contribution of alumni, to study student assessment, co-relation between different group of students.

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License.[5].Weka main interface is Explorer .Explorer provides Pre process panel to import data from database, CSV, Arff files etc.

The Classification panel enables the user to apply classification and regression algorithm to the data set, it also enable to generate tree, ROC curves etc. The Association rule panel provides algorithm to identify interrelationship among the attributes of the data. Clustering panel provides numerous algorithm for performing clustering on data set.

The Select attribute panel provides algorithm to identify most predictive attributes in the data set. The Visualize

panel provides a scatter plot matrix ,it can be selected ,enlarged and analysed .

II. RELATED WORK

Komal S. Sahedani ,Prof. B Supriya Reddy [6] in their finding they have shown that classification algorithms can be used successfully in order to predict a student's academic performance and to model the difference between fail and pass students. They also used feature selection technique to select the most affecting attributes in predicting students' dropouts.

Rakesh Kumar Arora and Dr. Dharmendra Badal [7] conducted study to predict the placement of the students by using data mining tool Weka .They applied the decision tree algorithm to identify those sets of students that are likely to face difficulty in getting the placements. The result of their will help the head of departments and placements in charge to design the strategies to improve the academic result, programming skills and communication skills of students.

Ramanathan.L, Swarnalatha P and D. Ganesh Gopal [8] used Sum of difference method to predict the student's placement by using some attributes such as academic records ,age and achievements etc. They have shown that SOD is admirable in terms of efficiency.

Samrat Singh and Dr.Vikesh Kumar [9] conducted study on engineering students to evaluate the performance of the students for the recruitment .They performed the comparative analysis of classification Data Mining algorithms using data mining tool Weka. The result of their finding shown that IB1 classifier is most suitable method for student's performance analysis.

III.METHODOLOGY

The method proposed in this paper for predicting student's placement belongs to the process of knowledge discovery and data mining(KDD). The main stages of method are:

A. Data Collection

In our study, we have collected data from final year students of computer engineering department of UCOE ,Patiala. During data collection some attributes such as Communication Skills, Programming Skills, Graduation result ,financial status etc are consider to predict student's placement .Table I shows the list of attributes about which data is collected from 122 students.

B. Data Pre-Processing

One of the important steps of Data Mining process is data pre-processing. Data Pre-processing is used in identifying the missing values, noisy data and irrelevant and redundant information from dataset. Before applying DM algorithm it is necessary to carry out some pre processing tasks such as cleaning ,integration and variable transformation. Firstly ,all the collected data is integrated into single datasheet of excel and after this we transform the continuous values into discrete values and numerical values into alphabetical values.

Then ,all the information is integrated in single datasheet and saved in CSV format.

C. Data Mining and Experimentation

In this research analysis of dataset is done with the help of WEKA 3.6 tool. it is one of the popular and open source

Attribute	Description	Possible values
Gender(gen)	Gender of the student	{male ,female}
Family income (fincome)	Family income/annual	{below middle class(bmc) if <= 4 lac, Middle class (mc) if >4lac<= 13lac, Above middle class (amc) if >13lac}
Categories(cat)	Student's category	{sc ,general, obc}
Location(loc)	Living loction	{rural(r),urban(u)}
Number of brother & sister (nsib)	Number of brother & sister of student	{yes-more than two, no-less than or equal to two}
Father occupation (fo)	Father occupation	{service, agriculture, business }
Mother occupation (mo)	Mother occupation	{service, businesswoman, homemaker, }
Hosteller(h)	Living in hostel	{yes(y),no(n)}
Tenth medium(tm)	10 th class medium	{Hindi, Punjabi ,English}
Twelfth result (twr)	12 th result	A-80-100 % B-70-89% C-69-60% D-59-33%
B.tech result(b.techr)	b.tech result	A-80-100 % B-70-89% C-69-60% D-59-33%
Programming skills(ps)	Programming skills	{good ,average ,poor}
Communication skills(cs)	Communication skills	{good ,average ,poor}
Aptitude skills(as)	Aptitude skills	{good ,average ,poor}

tool of data mining. It supports large number of data mining algorithms. Random tree search algorithm ,Id3

TABLE I :Attribute List

algorithm ,j48 algorithm of decision tree is used to classify or to generate rules from the student dataset of 120 students. These algorithms are applied with cross validation with 10 folds.

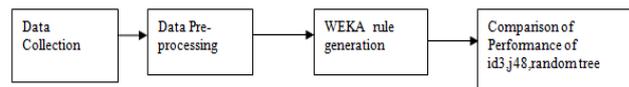


Figure 1.Methodology

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. [11] The decision tree algorithm is simple top

down greedy algorithm. The major step of algorithm is to continue to divide leaves that are not homogeneous into leaves that are as homogeneous as possible until no further division is possible. Figure 1 depicts the steps used to perform experimentation.

IV. RESULT AND DISCUSSION

The data set of 122 students is analyzed using ID3, J48 and Random decision tree algorithms.

The figure 2 shows the explorer interface of the WEKA tool. Explorer interface shows that 77 students got placed and 45 student are not placed from the dataset of 122 students.

The decision tree algorithms classified the student data set into two sets one is of placed student and other is of non placed students. Placed and non placed set is represented by "y" and "n" respectively.

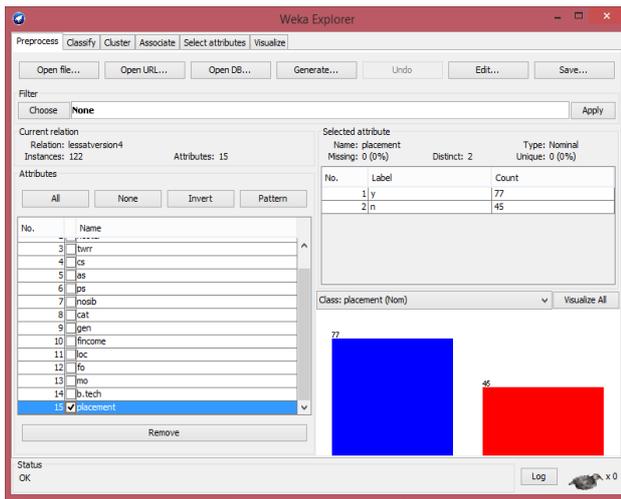


Figure 1. Student dataset open with WEKA Explorer

Table II compares the results of Random tree, ID3 and J48 decision tree algorithm with respect to different parameters.

Figure 3 represents the graph of table 2 and it shows that random tree correctly classified instances are more than ID3 and J48.

Table II : Result of Classifiers

Performance Parameters	Random Tree	ID3	J48
Correctly Classified Instances	109 (89.34%)	106 (86.88%)	107 (87.70%)
Incorrectly Classified Instances	13 (10.65)	14 (11.47%)	15 (12.29%)
Kappa Statistic	0.7701	0.7466	0.7322
Mean Absolute Error	0.1148	0.1167	0.1345
Root Mean Squared Error	0.3326	0.3416	0.3261
Relative Absolute Error	24.59%	25.33%	28.82%
Root Relative Error	68.89	71.08%	67.54%

Squared Error

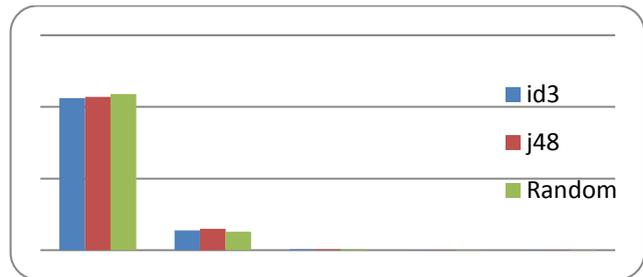


Figure 2. Comparison of Classifiers

Figure 4 shows the time taken by algorithms to build the model. The figure reveals that Random tree has taken only zero seconds to build the model. So it is little bit faster than ID3 and J48.

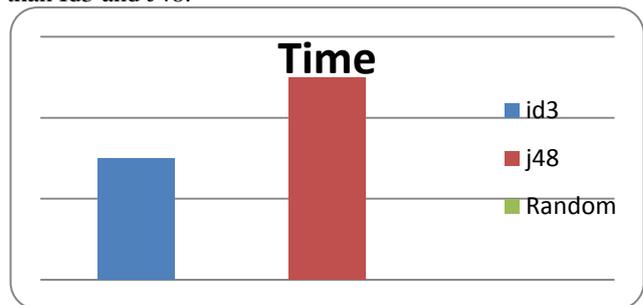


Figure 3. Time taken to build model

The figure 5 depicts that the accuracy of the Random tree is more as compared to ID3 and J48. The root absolute error (RAE) and root relative squared error (RRSE) are less as compared to other classifiers.

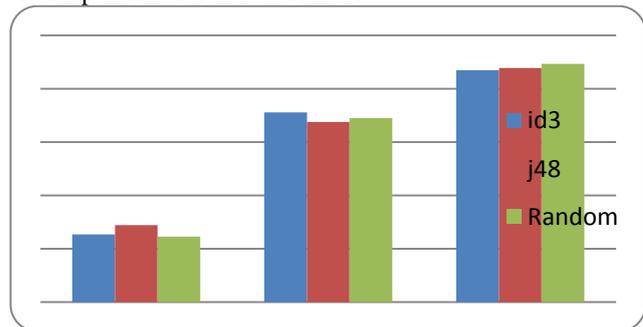


Figure 4: Accuracy

Figure 6 and Figure 7 show the accuracy of these classifiers for class Y and class N respectively.

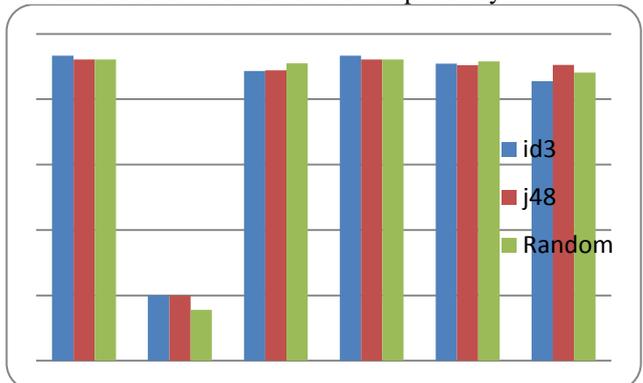


Figure 5. Accuracy For Class Y

The figure 6 depicts that TP rate of id3 seems little bit higher than J48 and Random for class "y". True positive rate(TP rate) is the ratio of positive instances that are correctly identified to the total number of actual positive instances. FP rate(false positive rate) is the ratio of negative instances that are incorrectly classified as positive to the total number of negative instances .Precision is the ratio of predicted correctly to the total predicted.

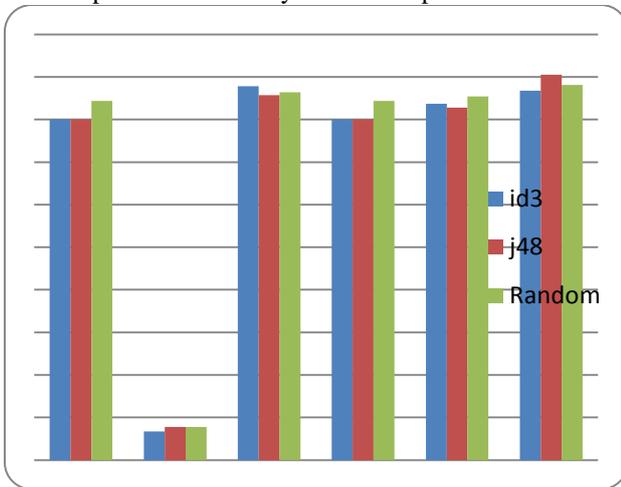


Figure 6.Accuracy for Class N

Confusion matrix for random tree is in table v shows the classification of 122 students dataset into two classes y is for placed student and n is for non placed student.

Table III: Confusion Matrix for Random Tree

a	b	<-- classified as
71	6	a=y
7	38	b=n

Table IV represents the Classification Model for Random tree algorithm. Which shows the classification of students into placed and not placed class on the basis various attributes.

TABLE IV: Classification Model for Random tree

```

RandomTree
fincome = amc
| b.tech = first : y (18/0)
| b.tech = d : n (3/0)
| b.tech = third
| | tenm = e : n (2/0)
| | tenm = h
| | | ps = good : y (1/0)
| | | ps = poor : y (0/0)
| | | ps = average
| | | as = good : y (0/0)
| | | as = poor : n (1/0)
| | | as = average : y (1/0)
| | tenm = p
| | | hostel = y : y (1/0)
| | | hostel = n
| | | | ps = good : y (1/0)
| | | | ps = poor : n (1/0)
| | | | ps = average : y (0/0)
| b.tech = second
| | cs = good : y (3/0)
| | cs = average
| | | ps = good : y (2/0)
    
```

```

| | | ps = poor : n (1/0)
| | | ps = average : y (3/0)
| | cs = poor : n (2/0)
fincome = mc
| twrr = first
| | b.tech = first : y (24/0)
| | b.tech = d : y (0/0)
| | b.tech = third : n (1/0)
| | b.tech = second : n (2/0)
| twrr = third
| | cs = good
| | | as = good : y (0/0)
| | | as = poor : n (1/0)
| | | as = average : y (3/0)
| | cs = average
| | | cat = general : y (1/0)
| | | cat = OBC : y (0/0)
| | | cat = SC : n (2/0)
| | cs = poor : n (4/0)
| twrr = second
| | gen = female
| | | ps = good : y (1/0)
| | | ps = poor : n (2/0)
| | | ps = average
| | | | cat = general
| | | | | nosib = yes : y (1/0)
| | | | | nosib = no
| | | | | b.tech = first : y (1/0)
| | | | | b.tech = d : y (0/0)
| | | | | b.tech = third : y (0/0)
| | | | | b.tech = second : n (1/0)
| | | | cat = OBC : n (1/0)
| | | | cat = SC : y (2/0)
| | gen = male
| | | as = good : y (3/0)
| | | as = poor : n (1/0)
| | | as = average : y (5/0)
| twrr = d
| | cs = good : y (1/0)
| | cs = average : n (1/0)
| | cs = poor : n (3/0)
| twrr = first : y (1/0)
fincome = bmc
| ps = good : n (1/0)
| ps = poor : n (9/0)
| ps = average
| | cs = good : y (2/0)
| | cs = average
| | | fo = service : y (1/0)
| | | fo = business : n (1/0)
| | | fo = agriculture : n (3/0)
| | cs = poor
| | | mo = service : y (1/0)
| | | mo = home-maker : n (2/0)
Size of the tree : 77
    
```

V. CONCLUSIONS AND FUTURE WORK

In this paper ,a simple decision tree methodology is used to predict the student placement and the comparative analysis of three decision tree algorithm that is id3,J48 and random tree is done. After Comparative analysis of these three algorithms shows that the performance of the random tree is better than the id3 and J48 algorithms in terms of accuracy and model build time. But the difference in the performance is not too big,So we

can say that any of these algorithms can be used for placement prediction. The generated decision rule can be used by the student to predict his/her campus placement. The result of these algorithms can be used by the placement in charge to identify the those set of students that are likely to face problem in campus placement. The classification model can play important role in increasing college placement. we also concluded that classification algorithms can be used successfully in order to predict student placement.

For the future work more attributes can be used to generate better student placement prediction model. The other data mining approaches such as association rule mining and clustering can be tested on student dataset for placement prediction.

ACKNOWLEDGMENT

I express my sincere gratitude to my Guide **Er.Kanwalpreet Singh Attwal** (Asst. prof) who assisted me throughout this work. I thank him for providing me there reinforcement, confidence and most importantly the track for this work whenever I needed it.

REFERENCES

- [1] R. Agrawal, "Mining Association Rules Between Sets of Items in Large Databases," in *ACM SIGMOD Conference on Management of data*, 1993.
- [2] S. B. Aher, "data mining in education system using WEKA," *International J. of Computer Applications*, pp. 20-25, 2011.
- [3] A. el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study," 2009.
- [4] C. B. e. a. Connolly T., "Database System: A practical approach to design, Implementation and management(3 rd edition)," Harlow; Addison-Wesley, 1999, p. 687.
- [5] [Online].
https://en.wikipedia.org/wiki/Weka_%28machine_learning%29.
- [6] P. B. S. R. Komal S. Sahedani, "Forecast Engineering Students Failure by Using Data Mining Techniques," *International Journal of Advance Engineer ing and Research Development (IJAERD)*, vol. 1, no. 5, 2014.
- [7] D. D. B. Rakesh Kumar Arora, "Placement Prediction through Data Mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 7, July 2014.
- [8] S. P. G. Ramanathan.L, "Mining Educational Data for students' placement prediction using Sum of difference method," *International Journal of Computer Applications (0975 – 8887)*, vol. 99, no. 18, 2014.
- [9] d. v. k. samrat singh, "performance analysis of engineering students for recruitment using classification data mining techniques," *IJCSET*, vol. 3, no. 2, pp. 31-37, 2013.
- [10] [online]http://www.estard.com/decisiontree/decision_trees_definition.asp .
- [11] [online] <http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf>.