iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59130

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Classification Model of Student Placement Datasets and Evolution of Classifiers of Weka Tool

Dr. Ramesh Prasad Aharwal
Assistant Professor Department of Mathematics Govt. P. G. College Damoh (M.P)

Abstract: Placement of students is one of the most important objective of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with .Placement is a foremost expectation concerning both the institute and the student's perspective. This paper analyze the different data mining techniques and implement data mining technique to extract Knowledge from MBA student placement datasets. In this regard we takes student placement data to extract pattern with using supervised learning techniques of data mining. In this paper we have used WEKA software for experimental work and author also try to evaluate some supervised learning classifiers of WEKA with using student placement dataset.

Keywords: Data Mining, Classification, J48, PART, Reptree, WEKA, Evolution measure parameters.

I. INTRODUCTION

Data Mining is a process of extracting previously unknown, valid, potential useful and hidden patterns from large data sets. As the amount of data stored in educational databases is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used [[9]. Classification is the most commonly applied data mining technique, which employs a set of pre-classified attributes to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms [1]. Data mining techniques are analytical tools that can be used to extract important knowledge from large data sets. The importance of data mining in higher educational institution proposing new techniques of data mining application in education like placement system and also focused on data mining capability to improve decision making processes in placement system in education institutions.

II. DATA MINING AND DATA MINING TECHNIQUES

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Various algorithms and techniques like Association rules, Classification, Clustering, Regression, Neural Networks, Fuzzy logic, Decision Trees, Genetic Algorithm, are used for knowledge discovery from databases Some are introduced as follows.

A. Classification

The most commonly useful data mining technique is classification which provides work for agroup of pre-classified data to develop a method that can classify the amount of large datasets. Applications like heart failure prediction, Student Placement prediction, fraud detection and many more.

B. PART Algorithm

Class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms. The main speciality of the PART algorithm is that it does not need to perform global optimisation like C4.5 and RIPPER to produce the appropriate rules.

1) J48 Algorithm

The C4.5 algorithm is implemented in WEKA tool as a classifier called J48 for building decision trees. C4.5 is basically an extension to the ID3 algorithm which was developed by Quinlan Ross.

It has the ability to not only manage the categorical attributes (as in ID3) but also continuous attributes. Gain Ratio, which is one of the attribute selection measures, is used by C4.5 to generate the decision tree. As per the algorithm, the gain ratio for each attribute is calculated and the attribute which gives the highest gain ratio is taken as the root node. Some unnecessary branches in the decision tree are also removed by C4.5, which is known as pruning to increase he accuracy of classification [9].

2) REPTree algorithm

REPTree is one of the fast decision tree classifier algorithms. It constructs the decision tree using entropy and information gain of the attribute with reduced error pruning technique. It constructs multiple trees and selects the best tree from the generated list of trees. REPTree prunes the tree using the back fitting method. REPTree algorithm sorts all numeric fields in the dataset only once at the start and then it utilize the sorted list to split the attributes at each tree node. It classifies the numeric attributes by minimizing total variance. The non-numeric attributes classified with regular decision tree with reduced error pruning technique.

III. RESEARCH OBJECTIVE

This study has aims to implement several prediction techniques in data mining to assist educational institutions with predicting their student's placement. If students are predicted to have low academic performance or less chance to get the placement, then extra efforts can be made to improve their academic performance and placement activity.

1) The main objective of this study is to use data mining methodologies to predict MBA student Placement based on student placement dataset. Data mining provides many tasks that could be used to study the recruitment of MBA students in various sectors.
2) To predict the placement or not placement class using classification algorithm.
3) To compare different classification algorithms.
4) To study the relationships among different factors deciding student placement in different sectors.

IV. RESEARCH METHODOLOGY AND EXPERIMENTAL SETUP

In the proposed study, WEKA 3.8 data mining tool is used to analyze and Extract Knowledge from MBA Student Placement datasets. There are a number of classification algorithms like Random forest, Random tree, j48, Naïve Bayes, REP Tree, SVM, and alike that can be used to construct models to predict the future data. In this study author select PART,J48 and Reptree to design a model for MBA student Placement dataset. Author have used Student placement dataset contains 215 instances and 13 Attributes. Description of datasets is presented in table 1.1

The various steps followed to build the prediction model and evolutions of models are:

Step1: Datasets are access from web resources.

Step2: Data pre-processing and it includes tasks like data cleaning, data integrating, data reduction and data transformation.

Step 4: Select the appropriate data mining technique according to the nature of problem and target variables.

Step 5: Apply the desired algorithm to the cleaned data set.

Step 6: Evaluate the performance of each classification algorithm and build the prediction model using the algorithm that out performs other classifiers.

Step 7: Result interpretation and conclusion

V. EVOLUTION MEASURES

A. FP Rate

FP rate is the proportion of all negatives that still produce positive test outputs. It can be calculated as:

$$\text{FP rate} = FP/FP+TN$$

B. Accuracy

Accuracy means the exactness of a predicted value to a known value. It can be calculated from this formula

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

C. Precision

Precision means positive predicted values, it can be calculated as:

$$\text{Precision} = TP/TP+FP$$

D. F-Measure

It measures the similarity of the known value and Prediction value distributions.it can be

$$\text{F-Measure} = 2(\text{Precision}*\text{Recall})/ (\text{Precision} +\text{Recall}$$

E. ROC Curve

An ROC curve is normally used way to visualize the performance of a classifier, and AUC is the best way to summarize its performance in a single number. An ROC curve plots True Positive Rate vs. False Positive Rate at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. From the figure we can interpret that the AUC (Area under the Curve) for these models is more than 0.75 or near about 0.75 . , which is good to be true.

F. Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

Table 1 WEKA generated Confusion Matrix of Classification Algorithms

PART	J48	Reptree
a b<-- classified as	a b<-- classified as	a b<-- classified as
43 4 | a = Placed	44 3 | a = Placed	45 2 | a = Placed
11 15 | b = Not Placed	10 16 | b = Not Placed	12 14 | b = Not Placed

VI. EXPERIMENTAL SETUP AND IMPLEMENTATION

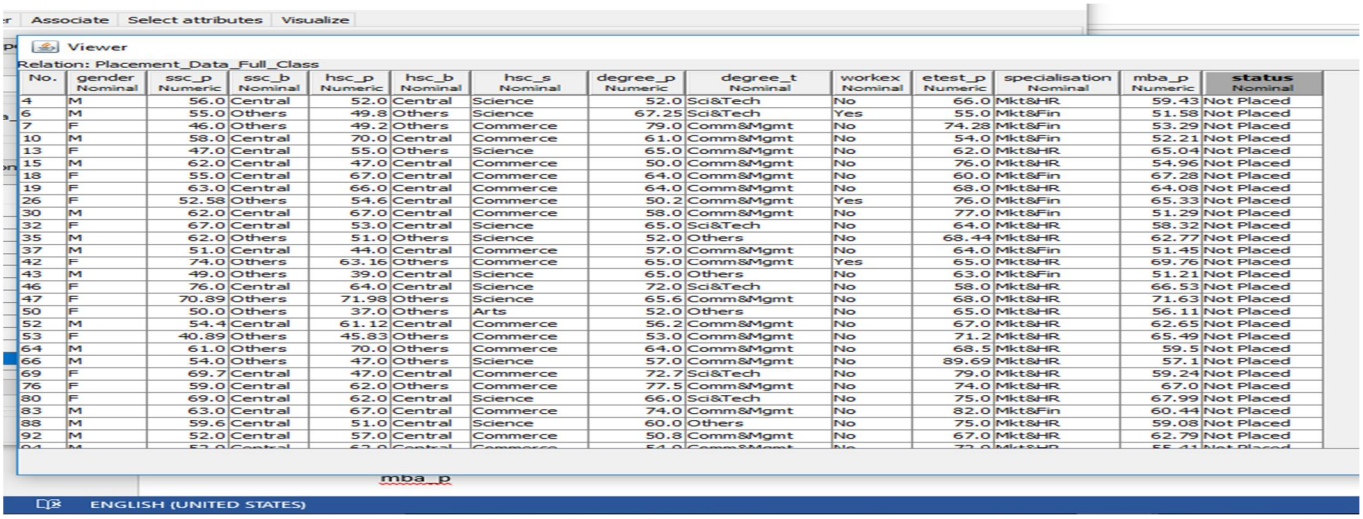
We have used WEKA tool for our implementation that has been developed at the University of Waikato in New Zealand. It comprises of an extensive collection of state-of-the-art machine learning and data mining algorithms which are written 9in Java. WEKA consists of all the tools that may be required to perform regression, classification, clustering, association rules etc. used datasets in this research is secondary data which is access from web resources. Three of the classification techniques that we have used and implemented on our dataset using Weka to build the classification models to predict the placement of the MBA students are: J48 decision tree algorithm, the PART and Reptree. Description of these algorithm have mentioned above. The results that were obtained after building the models are shown in the Result section.

A. Datasets and its Description

In this paper datasets was collected as a secondary data from web resources. Originally dataset is in csv (Comma Separated value) format. Collected data is required to preprocess.

Author have used MBA Student placement dataset which have taken from web resources. Dataset contains 215 instances and 13 Attributes. Name of attributes are Gender , ssc_p , ssc_b, hsc_p, hsc_b, hsc_s, degree_p, degree_t, workex, etest_p , specialization , mba_p, status (Class Attribute). Here I take Status attribute as a class attribute. Description of datasets is presented below.

Table 2 Datasets Description



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

VII. RESULT AND DISCUSSION

The experiment was conducted by using student placement datasets which have 215 instances and 13 variables. To predict the MBA student placement based on variable which are shown in above mentioned table. Experiment have done the classification algorithms like J48, PART and Reptree are used. Datasets is divided into two subsets such as 66% for training set and remaining for as a test dataset. these subsets are used for training as well as testing which gives better result than other testing options. By considering result as the target attribute in such a manner is split 66.0% as a train, remainder test done on data set. Author observe that J48 model is better than others. Result shown in figure 3 and table 3. Prediction model are depicted in fig. 4 and table 4.

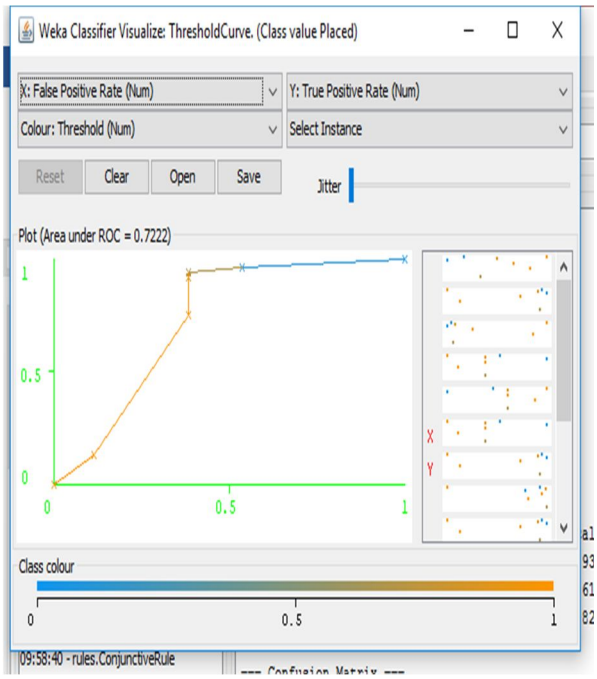


Fig. 1 ROC curve for Placed class of J48

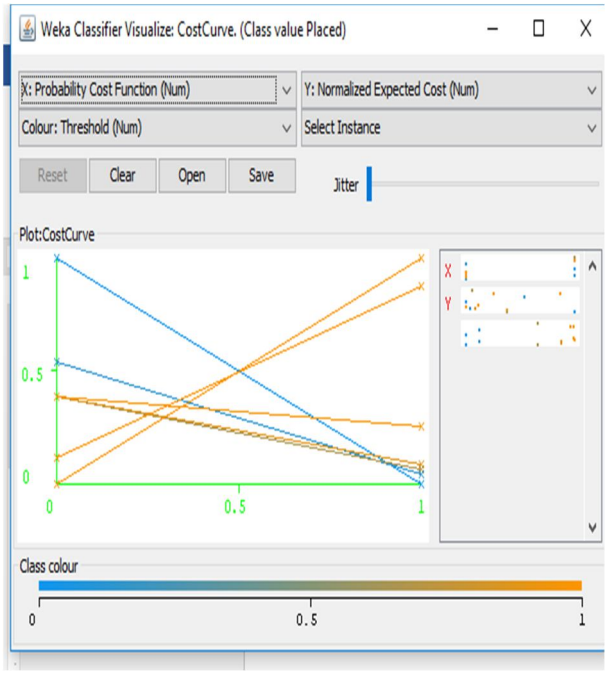


Fig. 2 Cost curve for Placed class of J48

Table 3 Model evolution parameter values which find during experimental work

Fig. 3 Evolution of Models

Evolution measures for Placed Class	PART	J48	RepTree
TP Rate	0.915	0.936	0.957
FP Rate	0.423	0.385	0.462
Precision	0.796	0.815	0.789
Recall	0.915	0.936	0.957
F-Measure	0.851	0.871	0.865
ROC Area	0.798	0.722	0.74
Correctly Classified %	79.4521 %	82.1918 %	80.8219 %

B. PART Generated Rules

Table 4 Learning rules or Pattern from PART

hsc_p<= 55 AND ssc_p<= 71: Not Placed (27.0	workex = Yes AND ssc_p> 52: Placed (24.0/1.0)
ssc_p> 64 AND specialisation = Mkt&Fin AND hsc_b = Others AND ssc_p> 70.5: Placed (42.0)	workex = No AND degree_t = Comm&Mgmt AND specialisation = Mkt&HR AND ssc_p<= 71 AND etest_p> 58 AND mba_p> 56.49: Not Placed (10.0)
ssc_p<= 55 AND hsc_s = Commerce: Not Placed (9.0	workex = No AND degree_t = Sci&Tech AND hsc_p<= 65.5: Not Placed (7.0)
degree_p> 65.6 AND gender = M: Placed (44.0/2.0)	workex = No AND degree_t = Comm&Mgmt AND specialisation = Mkt&Fin AND degree_p> 61: Placed (8.0/1.0)
workex = No AND hsc_s = Commerce AND specialisation = Mkt&Fin: Not Placed (6.0/1.0)	

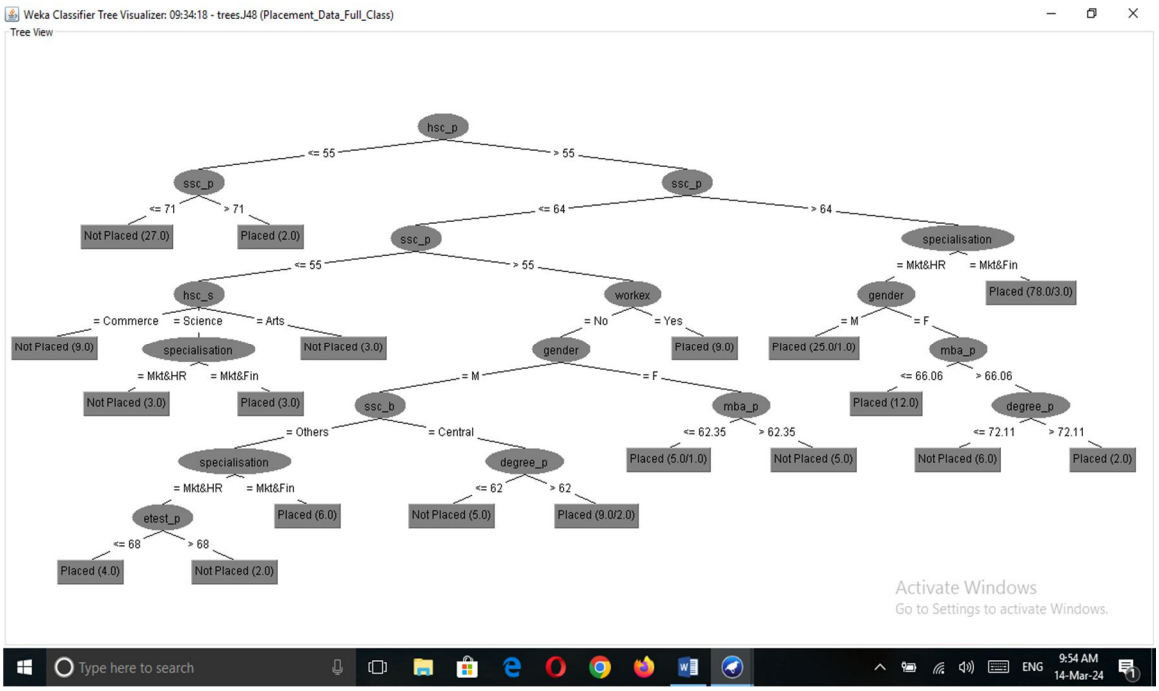


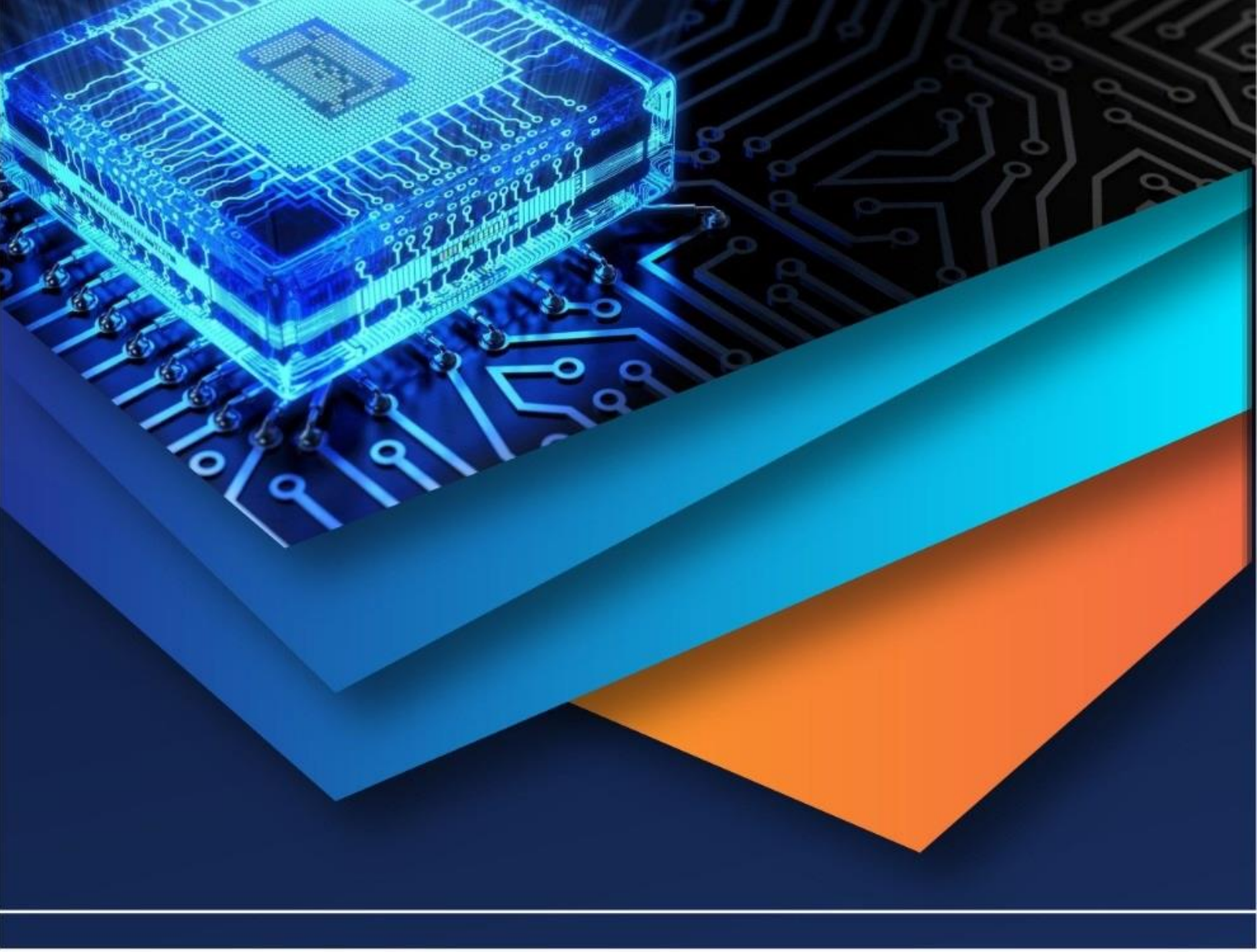
Fig. 4 Decision Tree Generated from J48 of WEKA

VIII. CONCLUSION

The performance measurement of the model was evaluated with the help of various metrics like accuracy, sensitivity, F1-score and precision. The performance visualization of the binary class classification problem was analyzed using a graphical plot AUC (Area under the Curve) ROC (Receiver Operating Characteristics) curve and Cost Curve. The ROC curve is generated by plotting the true positive rate against false-positive rates at various threshold rates. The best algorithm based on the performance parameters was selected to predict the placement category of MBA students. The ROC Curve and COST Curve of J48 Models for Placed Class are shown in Fig. 1 and Fig. 2. Predictive model result are shown in the Table 1, fig. 4., Table 4.

REFERENCES

[1] Dey, K. Abhirup, and A. Kumar, Prediction and Analysis of Student Performance by Data Mining in WEKA. 2018.
[2] A. Kumar Pal and S. Pal, "Classification Model of Prediction for Placement of Students," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 11, pp. 49–56, 2013, doi: 10.5815/ijmecs.2013.11.07.
[3] A. S. Rao, S. V. Aruna Kumar, P. Jogi, K. Chinthan Bhat, B. Kuladeep Kumar, and P. Gouda, "Student placement prediction model: A data mining perspective for outcome-based education system," Int. J. Recent Technol. Eng., vol. 8, no. 3, pp. 2497–2507, 2019, doi: 10.35940/ijrte.C4710.098319.
[4] D. Manjusha, B. Pooja, A. Usha, and B. E. Scholars, "Student Placement Chance," vol. 7, no. 5, pp. 1011–1015, 2020.
[5] G. B. Tarekegn, "Application of Data Mining Techniques to Predict Students Placement in to Departments," Int. J. Res. Stud. Comput. Sci. Eng., vol. 3, no. 2, pp. 10–14, 2016, doi: 10.20431/2349-4859.0302002.
[6] M. Kumar and A. J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 8, pp. 25–31, 2017, doi: 10.5815/ijmecs.2017.08.04
[7] P. Manvitha and N. Swaroopa, "Campus Placement Prediction Using Supervised Machine Learning Techniques," Int. J. Appl. Eng. Res., vol. 14, no. 9, pp. 2188–2191, 2019, [Online]. Available: http://www.ripublication.com.
[8] R. S. Kumar, F. Dilsha, A. N. Shilpa, and A. A. Sumayya, "Student Placement Prediction Using Support Vector machine Algorithm," vol. 9, no. 5, pp. 40–43, 2021, doi: 10.17148/IJIREEICE.2021.9507.
[9] S. Kalaivani, B. Priyadharshini, and B. S. Nalini, "Analyzing Student's Academic Performance Based on Data Mining Approach," Int. J. Innov. Res. Comput. Sci. Technol., vol. 5, no. 1, pp. 194–197, 2017, doi: 10.21276/ijircst.2017.5.1.4.
[10] S. Samrat . K. Vikesh "Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques", IJCSET |February 2013 | Vol 3, Issue 2, 31-37 ,2013
[11] S. Sajwan, R. Bhardwaj, R. Mishra, and S. Jaiswal, "Student Placement Prediction Using Machine Learning Algorithms," Lect. Notes Electr. Eng., vol. 1061 LNEE, pp. 231–241, 2024, doi: 10.1007/978-981-99-4362-3_22.
[12] Shruthi P, "Student Performance Prediction in Education Sector Using Data Mining," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 6, no. 3, p. 2277, 2016, [Online]. Available: www.ijarcsse.com.
[13] T. Patel and A. Tamrakar, "a Data Mining Techniques for Campus Placement Prediction in Higher Education," Indian J.Sci.Res, vol. 14, no. 2, pp. 467–471, 2017.
[14] T. Joseph, "Placement Prediction of Students Using the Data Mining Tool Weka," vol. 3, no. 1, pp. 9–11, 2021, doi: 10.5281/zenodo.5093609.
[15] Y. Ingale, T. Bedse, S. Khairnar, and D. Ghute, "Student's Placement Prediction Using Support Vector Machine," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., no. May, pp. 55–60, 2020, doi: 10.32628/cseit20651



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY