International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.06, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A Classification Model for Predicting Campus Placement performance Class using Data Mining Technique.

K.Manikandan[1], S.Sivakumar[2], M.Ashokvel[3]

[1]*Assistant Professor, Department of Computer Applications,A.V.C. College of Engineering,Mayiladuthurai, Tamilnadu, India.*
[2]*Assistant Professor, Department of Computer Applications,A.V.C. College of Engineering,Mayiladuthurai, Tamilnadu, India.*
[3]*Assistant Professor, Department of Computer Applications,A.V.C. College of Engineering,Mayiladuthurai, Tamilnadu, India.*

ABSTRACT

Data Mining is very useful in the field of education to predict campus placement performance of the students. Placement of students is one of the important activities in educational institutions. Admission and reputation of institutions mainly depends on placements. The overall goal of data mining is to extract information from a dataset and transform it into useful structure for further use. This can help in building new systems and take decision making in educational system. This paper discusses use of classification algorithms in educational data mining. Classification algorithms is applied on student previous year data to compute the distribution of placement class and it can be used to predict the students' placement in various companies. This will help the students to identify the category of company in which they are eligible and prepare accordingly in an efficient manner.

Keywords —Data Mining, Classification, Naive Bayesian, placement Prediction

I .INTRODUCTION

Data mining is the process of discovering interesting knowledge from large amount of data stored in database, database warehouse or other information responsibility. Data mining term is mainly used for the specific of six activities namely Classification, Estimation, Prediction, Association rules, Clustering, Description and Visualization. The first three tasks – classification, estimation and prediction are all examples of directed data mining or supervised learning. Majority of students in higher education join a course for securing a good job. Therefore taking a wise career decision regarding the placement after completing a particular course is crucial in a student's life. Data mining is one of the important techniques used in Education field. In real world, predicting the performance of the student's placement is a challenging task. The primary goals of Data Mining in practice tend to be Prediction and Description. Predicting performance involves variables like X Mark, XII Mark, UG Mark, PG Mark Programming language, etc. in the student database to predict the unknown or future values of interest. Educational Data Mining uses many techniques such as Decision Trees, Multilayer Perception, NaïveBayes and many others. Using these methods many kinds of knowledge can be discovered. The aim of classification is to predict the future output based on the available data. The prediction of computer science students where they can be placed after the completion of their course will help to improve efforts of students for proper progress. It will also help teachers to take proper attention towards the progress of the student during

the course. It will help to build reputation of institute in existing similar category institutes in the field of IT education. The present study concentrates on the prediction of placements of Computer Science students. We apply data mining techniques using J48 Algorithm, REPTREE Algorithm, Naïve Bayes classifier,BayesNet classifier, Multilayer Perceptron to interpret potential and useful knowledge.

II EDUCATIONAL DATA MINING

Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students performance. Different from data mining methods, EDM, when used explicitly, accounts for (and avail of opportunities to exploit) the multilevel hierarchy and lacks independent educational data.Educational data mining methods come from different literature sources including data mining, machine learning, psychometrics, and other areas of computational modelling, statistics, and information visualization.

III CLASSIFICATION ALGORITHMS

Educational Data mining can be implemented in many techniques such as decision trees, neural networks, k-nearest Neighbor, Naïve Bayes, support vector machines and many others. Using these methods many kind of knowledge can be discovered such as association rules, classification, clustering, and pruning the data. The classifiers used in this paper consists of common decision tree algorithm C4.5 (J48), REPTREE Algorithm, Naive Bayes classifiers, BayesNet classifiers and Multilayer Perceptron (MLP) algorithm. The results obtained from the classification task are presented in the experimental results.

A. J48 Algorithm

The J48 classification algorithm is WEKA's version of the implementation of the C4.5 decision tree algorithm, which uses a greedy technique to induce decision trees and make use of reduced- error pruning. The algorithm was developed from ID3 algorithm for handling missing data, continuous data, pruning, splitting and generating rules [12]. The technique uses Gain Ratio instead of Information Gain for splitting purpose:

$$\text{Gain Ratio (D, S)} = \text{Gain (D, S)/ Split INFO}$$

$$\text{Where, Split INFO} = -\left(\sum_{i=1}^{s} \frac{D_i}{D} \log_2 \frac{D_i}{D}\right)$$

In order to categorize a given set, Information Gain as a metric is compulsory, with a function to deliver a balance inthe splitting. Providing a data set that containsattributes, we can measure the entropy as a degree ofimpurity

$$Entropy = \sum_j -P_j \log_2 P_j$$

And determining the best attribute for a node in the tree, we use the Information Gain as a measure, such that Information Gain, Gain (S,A) attributes are defines as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

B. REPTREE Algorithm

The REPTREE classification algorithm is a technique that builds trees using entropy as impurity measure and also makes use of reduced-error pruning.

C. NAIVE BAYES Algorithm

The algorithm is based on Bayes rule of provisional possibility and adopts independence between attributes values in a data set [14]. The algorithm requires small amount of training data to predict a classification model. The technique signifies a method to probabilistic discovery of knowledge and gives efficient algorithm for data classification [15]. The algorithm makes use of the Bayesian theorem with naïve independent assumptions as in the formula [3]

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

D. BayesNet Classifiers

A Bayesian network, Bayes network model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).

E. Multilayer Perceptron

Multilayer Perceptron (MLP) algorithm is one of the most widely used and popular neural networks. The network consists of a set of sensory elements that make up the input layer, one or more hidden layers of processing elements, and the output layer of the processing elements. MLP is especially suitable for approximating a classification function (when we are not so much familiar with the relationship between input and output attributes) which sets the example determined by the vector attribute values into one or more classes.

IV DATA COLLECTION

Table 1: The definition of attributes and values

Attribute	Definition	Values
Roll number	Identity of a student	ROLL NO
SSLC	X Standard board marks	1 – 100
HSC	Higher secondary board marks	1 – 100
UG	Under graduate marks	1 – 100
PG	Post graduate marks	1 – 10
Specialization	Programming Knowledge	Java, .Net
Company Level	Company Details about selection in campus interview	T1 , T2, T3

In this case, we have collected 2013-2016 batch students' details as a dataset from computer application department in A.V.C. College of Engineering. Here, Company Status T1 refers to Tier1 level company, T2 refers to Tier2 level company and T3 refers to Tier3 level company.

V DATA PREPROCESSING

Raw data is a quality less and inconvenience data for processing. This poor quality of raw data affects the data mining efficiency. In order to improve the quality of the data and, also the mining results pre-processing of raw data is carried out. Data preparation and filtering steps takes large amount of time. In this case, we have collected computer application students details as a dataset. Here, irrelevant attributes such as students residential address, name, etc had been removed.

VI IMPLEMENTATION OF MINING MODEL

Weka is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. From the above data, placement.arff file was created. This file was loaded into WEKA explorer. The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. The algorithm used for classification is J48 Algorithm, Naive Bayes, BayesNet, REPTree and Multilayer Perceptron (MLP). Under the "Test options", the 10-fold cross-validation and percentage split are selected as our evaluation approach. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. This predictive model provides way to predict whether a new student will place or not in an organization.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.06, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

VII EXPERIMENTAL RESULTS AND DISCUSSIONS

The objective of the study is to explore if it is possible to predict the student placement based on the various input variables which are retained in the model. The classification model was built using several different algorithms and each of them using different classification techniques. The WEKA Explorer application is used at this stage. Each classifier is applied for two testing options – cross validation and percentage split. The screen shot of the WEKA preprocessing stage is shown in Figure 1.

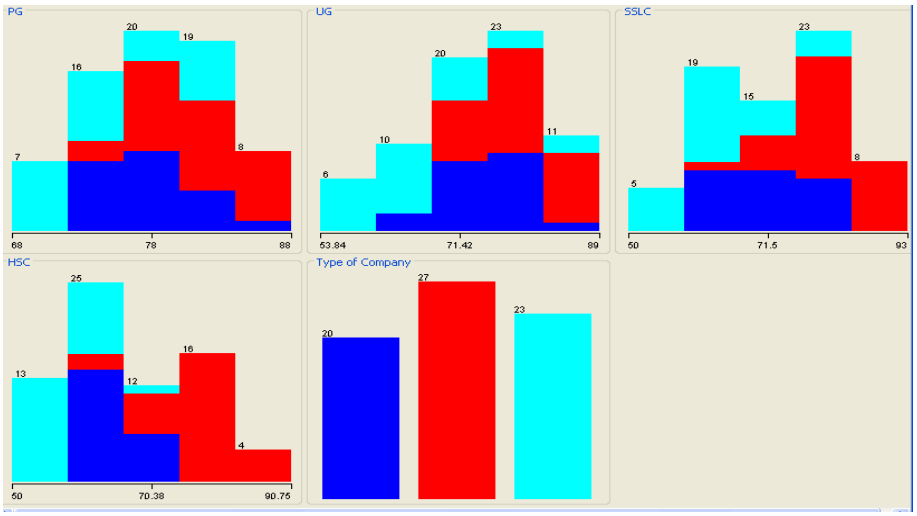

Figure 1. WEKA Screenshot of Data Distribution in the Preprocessing Stage.

Table 2. Classification results for the decision tree algorithm

	J48 – 10-fold Cross validation		J48 – Percentage split	
Class	**TP Rate**	**Precision**	**TP Rate**	**Precision**
T1	0.889	0.857	0.909	1
T2	0.85	0.68	0.75	0.375
T3	0.652	0.882	0.556	0.833
Weighted Avg.	0.8	0.815	0.75	0.833

7.1 Results of Decision Tree Classifier

In this study, J48 classification algorithm was implemented on the data and the results of the classification is presented in Table 2. It is inferred from the Table 2, that J48 has correctly classified about 80% for the 10-fold cross-validation testing and 75%for the percentage split testing. It produces a classification tree. The screenshot of decision screen building process in shown in Figure 2. The results from Table 2 reveal that the True Positive Rate is high for the classes – T1 (90 %) and T2 (75-85 %).

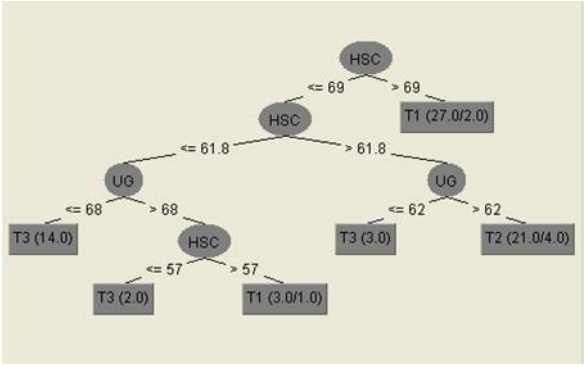

Figure 2. Screenshot of Decision Tree build using J48 Classifier.

Table 3. Classification results for the REPTreeAlgorithm

	REPTree– 10-fold Cross validation		REPTree– Percentage split	
Class	**TP Rate**	**Precision**	**TP Rate**	**Precision**
T1	0.852	0.92	0.73	0.75
T2	0.65	0.52	0.75	0.188
T3	0.652	0.75	0.444	1
Weighted Avg.	0.729	0.75	0.417	0.75

7.2 Results of REPTreeClassifier

Table 3 shows the classification results for REPTreeclassifier. The REPTreeclassifier correctlyclassifies about 72.9% for the 10-fold cross-validation testing and 41.7 % for thepercentage split testing. The results from Table 3 show that the True PositiveRate is high for the T1 (73-85%) and T2(65-75%).The precision is found to be high for theclasses – T1 and T3.

Table 4. Classification results for the Naive Bayes Classifiers

	Naive Bayes– 10-fold Cross validation		Naive Bayes– Percentage split	
Class	**TP Rate**	**Precision**	**TP Rate**	**Precision**
T1	0.852	0.821	0.909	1
T2	0.65	0.722	0.75	0.6
T3	0.87	0.833	0.889	0.889
Weighted Avg.	0.8	0. 797	0.875	0.892

Table 5. Classification results for the BayesNet Classifiers

	BayesNet– 10-fold Cross validation		BayesNet– Percentage split	
Class	**TP Rate**	**Precision**	**TP Rate**	**Precision**
T1	0.889	0.923	0.909	0.909
T2	0.7	0.667	0.75	0.429
T3	0.783	0.783	0.667	1
Weighted Avg.	0.8	0.804	0.792	0.863

7.3 Results of Bayesian Classifiers

The present study implements Bayesian classifiers namely Bayesian networks and naive Bayes on the dataset and the results are presented in Table 4 and Table 5. Table 4presents the classification results for Naive Bayes classifier and it is found that Naive Bayes classifier correctly classifies about 80 % for the 10-fold cross-validation testing and 87.5% for the percentage split testing.The results from Table 3 reveal that the True Positive is high for most of the classes – T3, T1 and T2. The precision is also high for the classes –T1, T3 and T2. Table 5 presents results of BayesNet classifier on the dataset. It can be verified that Bayes Net correctly classifies about 80 % for the 10-fold cross-validation testing and 79.2 % for the percentage split testing.The results from Table 5, shows that the True Positive Rate is high for the classes –T1, T3 and T2. The precision is also high for the classes – T1.

Table 6. Classification results for the MLP Classifiers

	MLP – 10-fold Cross validation		MLP – Percentage split	
Class	**TP Rate**	**Precision**	**TP Rate**	**Precision**
T1	0.852	0.958	0.818	0.9
T2	0.65	0.619	0.75	0.5
T3	0.826	0.76	0.889	1
Weighted Avg.	0.786	0.796	0.833	0.871

7.4 Results of Multilayer Perceptron Classifier

The present study implements MLP on the dataset and the results are presented in Table 6. It presents the classification results for MLP classifier correctly classifies about 78.6 % for the 10-fold cross-validation testing and 83.3% for the percentage split testing. The True Positive is high for most of the classes – T1 and T3. The precision is also high for the classes – T1 and T3.

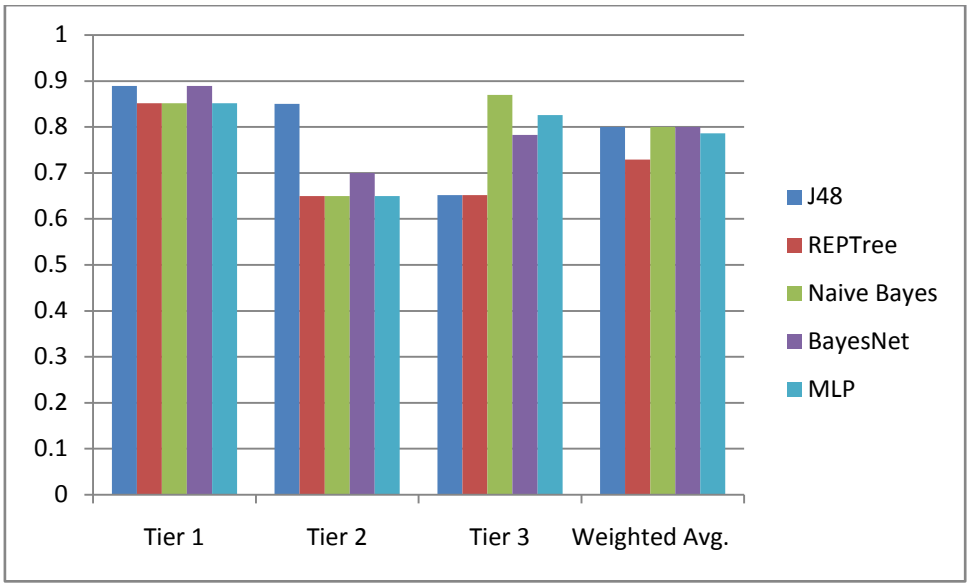

Figure 3. Classification algorithms performance comparison.

The results of the classification reveals that the Bayesian classifiers performs very well in comparison with other classifiers with the highest overall accuracy, followed by J48 Classifiers. REPTree classifier and MLP Classifier performs poorly and are less accurate than the others. The overall accuracy of all the tested classifiers is well above 70%. Naïve Bayes and BayesNet registered accuracy for 80 %. J48 produces accuracy very near to 80 %. In addition, further detailed analysis of the classification accuracy for the different classes reveals that the predictions are worst for the Tier 2 class. The classification accuracy is very good for Tier 1 class.

VIII CONCLUSION

The J48, Naive Bayes, BayesNet, REPTree and Multilayer Perceptron (MLP)data mining techniques were implemented on students data for analyzing the students placement selection..The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform among the algorithms. The range of prediction varies from 72-80 %. Bayesian classifiers perform very well in comparison with other classifiers with the highest overall accuracy. Moreover, the classifiers perform differently for the three classes. The data attributes that are found to have significantly influenced the classification process are Tier 1 and Tier 3 classes. The study can be further extended to performance of other classification techniques with larger sample dataset.

REFERENCES

[1] David Hand, HeikkiMannila, Padhraic Smyth Principles of Data Mining

[2] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," Procedia-Social and Behavioral Sciences vol. 97, pp. 320–324, 2013

[3] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker, Handbook of educational data mining. CRC Press, 2010

[4] Badr, E. Din, and I. S. Elaraby, "Data Mining : A prediction for Student ' s Performance Using Classification Method," World J. Comput. Appl. Technol., vol. 2, no. 2, pp. 43–47, 2014.

[5] K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification," Int.J. Comput. Sci. Inf. Secur., vol. 9, no. 4, 2011.

[6] E. Review, "Data mining approach for predicting student performance," J. Econ. Bus., vol. X, no. 1, pp. 3–12, 2012.

[7] P. Ajith, B. Tejaswi, and M. S. S. Sai, "Rule Mining Framework for Students Performance Evaluation," Int. J. Soft Comput. Eng., vol. 2, no. 6, pp. 201–206, 2013.

[8] K. Daimi and R. Miller, "Analyzing Student Retention with Data Mining.," in International Conference on Data Mining, 2009, pp. 55–60.

[9] Baker RSJD. Data mining for education. International encyclopedia of education. 2010; 7:112–8.

[10] Pal AK, Pal S. Analysis and Mining of Educational Data for Predicting the performance of Students. International Journal of Electronics Communication and Computer Engineering. 2013; . 4(5):1560–5.

[11] Rathee A, Mathur RP. Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance. International Journal of computers & Technology. 2013;4(2):244–7.

[12] Aher SB, Lobo LMRJ. Data mining in educational system using Weka. IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT). 2011; 3:20–5.

[13] Ajith P,,Tejaswi B, Sai MSS. Rule Mining Framework for Students Performance Evaluation. International Journal of Soft Computing and Engineering. 2013; 2(6):201–6.

[14] Anuradha., C and T. Velmurugan, A Comparative Analysis on the Evaluationof Classification Algorithms in thePrediction of Students Performance, Indian Journal of Science and Technology, Vol 8(15), July 2015

[15] Samrat Singh, Dr. Vikesh Kumar" Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education", International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4, August 2012.

[16] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of Students", I.J.Modern Education and Computer Science, 2013, 11, 4956Published Online November 2013 in MECS.

[17] NeelamNaik, SeemaPurohit**"**Prediction of Final Result and Placement of Students using Classification Algorithm**"** International Journal of Computer Applications (0975 – 8887) Volume 56– No.12, October 2012

[18] V.Ramesh, P.Parkavi, P.Yasodha" Performance Analysis of Data Mining Techniques or Placement Chance Prediction" International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011 1 ISSN 2229

[19] A. K. Pal, and S. Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students", (IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.

[20] B.K. Bharadwaj and S. Pal., *Data Mining: A prediction for performance improvement using classification",* International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.