International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 739-745 June 2018-Dec 2018 www.csjournals.com

Critical Evaluation of Classification Algorithms for Performance Prediction in Higher Education Setup

Tripti Mishra[1], Dr. C.D. Kumawat[2]
Research Scholar[1], Professor[2]
Department of Computer Science, Mewar University, Rajasthan

Abstract: Data Mining has emerged as an interesting research field which has found wide application in multiple areas like agriculture, medicine, marketing etc. Various functionalities of data mining like Clustering, Classification, Outlier detection, Discrimination has given new direction to descriptive and predictive research. Classification algorithms have been used extensively for predicting the category of a dependent variable based on independent variables. Various classification algorithms like Decision Tree, Naïve Bays, Support Vector Machines and Ensemble Methods have been developed, but for a given problem it is essential to scrutinize the performance of the algorithms and then select the best one for carrying out prediction.
This paper presents a comparative evaluation of four algorithms J48 (Decision Tree), Naïve Bays , and Random Tree, Random Forest (Ensemble Methods) in predicting the academic performance of the students of Master of Computer Application Course (MCA) affiliated to Guru Gobind Singh Indraprastha University, Delhi. The comparison is done on the performance parameters like correctly classified instances,precision, recall and time taken. J48 is found to be most suitable Algorithm for performance prediction.
Keywords: EducationalData Mining, Classification Algorithms, Performance prediction.

1 INTRODUCTION

Data Mining focuses on discovering interesting and useful patterns hidden in the data generated in various fields like Banking, Agriculture, E-commerce and Medicine etc. Educational data Mining has come up as a new and exciting research area where various classification algorithms are being used for predicting students result. Classification algorithms is also known as supervised learning as the data already indicate few fixed number of categories in which the predicted variable will fall based on independent variable. There are two phases of each classification algorithm. In the first phase, a training dataset consisting of a set of independent attributes and one target (dependent) attribute is considered and algorithms are used to derive a model for determining the class label of target attribute based on independent attributes. Once the model is finalized it is validated, using a testing data set, that is the remaining and determining the accuracy of the model. If the model accuracy is acceptable the model can be applied to predict the class label of an unseen data whose class label is not known. There is no thumb rule for suitability of algorithm to any given problem and various algorithms are applied and evaluated on performance parameters like accuracy, TP rate, Precision, recall and time taken to build the model.

Accuracy is fraction of correctly classified records from all the records.Precision(p) determines the fraction of records that turns out to be positive from that have been declared positive by the classification.Recall (r) measures the fraction of total positive records that have been classified as positive, which is like True Positive Rate. Interpretability is nowadays considered as a major criteria for selecting the classifier specially where it is important to identify which are the independent factors that are affecting the target variable. This paper is an attempt to compares four algorithms J48, random Tree, random forest and BayesianNetwork for the problem of academic performance prediction.The paper has been organized

The next section presents Review of Literature on Classification algorithms,Research methodology is discussed in Section 3, Section 4gives the analysis of experiment, followed by Result and Discussions in section 5 and the section 6 concludes the paper.

2 LITERATURE REVIEW

Performance prediction is most actively explored field of educational data mining.The literature is reviewed in chronological order to identify the algorithms used by researchers.

In their research work[7], authors found thatthe decision tree was more accurate than the Bayesian network in predicting result of final year post graduate students of Thao University of Vietnam and Asian Institute of Technology.

Various types of decision tree like Chi-Square Adjusted Interaction Detection(CHAID), C5, Classification And Regression Trees (CHAID) and Artificial Neural Network ANN were evaluated by [9] and it was shown that C4.5 exhibited highest accuracy (97.3%) followed by CART (92.5%), ANN (91.42%) CHAID (57.77%) for the development of a student performance assessment and monitoring system.

Academic performance of senior secondary school was predicted by [10] using attributes like University matriculation exam, GCE (General Certificate of Education) Score, Senior Secondary Certified Examination score SSCE), grades in O level subject, location of University from, gender and age, Cumulative Grade point was to be classified as Good, Average and Poor. ANN was able to give 74% accuracy of prediction. Instead of Higher education [4] have used Decision tree, Neural network, Support Vector machine Random forest to predict the academic performance of senior secondary students and concluded that attendance, parents job, previous year performance affect the current performance.[11] has experimented secondary student from different schools of Tamilnadu and observed that classification methods like Naïve Bayes, one R voted perception performed much better with feature selected subset than where all variables were considered.[1] have compared C4.5,NB Tree, Bayes Net, Hidden Naïve Bayes , and voting techniques of classification based on three weak classifiers (Naïve Bayes, One R and Decision Stump) for improving the accuracy of performance prediction. The combination of HNB method and one weak classifier Decision Stump uses for voting technique. This particular combination has been taken as HNB works well with most of the classes except for a high distribution class where Decision Stump gives good result.The students dropping out of an open polytechnic of New Zealand due to failure has been explored by[7].CART algorithm was found to be best suited with highest accuracy.

Expanding their earlier work, [7] applied classification on data of students enrolled in Open Polytechnic programme. Ethnicity, Course level, Secondary school, age, Course block , Course offer type and Work status have been found most important attributes after feature selection. The logistic regression and discriminant analysis models were found to be a superior prediction model with higher accuracy than the classification tree models (between 1% to 4%) however at the cost of using more variables. The authors therefore recommend the use of the CART classification tree model in the early identification of at risk students.

SQL server and Analysis services was used to construct performance meodel of Mugla University, Indonasia by[5]. The study aims to discover individual characteristics that decide their success using Microsoft Decision Tree. Ranking of attributes was done by [1]. Their research emphasized that Naïve Bayes gave highest accuracy followedby Logistic and decision tree.

[13] applied smooth support vector machine (SSVM) classification and Kernal K means clustering techniques to develop a model of student academic predictors by employing psychometric factors such as Interest, Study behavior Engage Time, and Family Support. Successful portioning was obtained with 5

clusters J48 decision tree was used to generate predictive rules which was implemented into SSVM algorithm to predict the student final grade.[2] based their experiment only on Previous Semester marks, class test grade, seminar performance, Assignment, attendance, Lab work to predict end semester marks.

[12] recommends use of SVM when individual academic performance of the student is to be predicted while multiple regression technique is best suited when average performance of whole class is to be predicted.

[3] concluded that Decision tree proves to be a better classifier than the neural network with 1.31% more accuracy.[14] have ranked importance of 24 predictor variables including demography, scores in Maths, Turkish, religion and ethics, science and technology and level determination exams etc for predicting Turkish secondary education placement result. Application of Artificial Neural Network, Support Vector Machine, Multiple Regression and Decision Tree were considered and decision Tree C5 gives best result.

In [16] authors have applied Chi – square test, One-R test, Information Gain test and Gain – ratio test and affirmed that GPA, score of entrance exam, study material and average weekly hours devoted to studying are having maximum impact while number of household member distance of residence and gender have least impact.[15] applied various algorithms decisions tree (C45 Random Forest, BF Tree, Rep Tree)Functions (logistic RBF Network) Rule (3 Rip) and Bayes Net, Naive Bayes to categorize (predict) students in 5 categories (Very good, Good, Satisfactory, Below Satisfactory and Fail)Random Forest proven to most accurate classifier .

[6] used 14 attributes including personal profile, secondary educational score, entrance exam score, admission year and used l the classifier J48, Bayesian, K-nearest neighbor one R and J Rip. In [6] Kabakchieva concludes that J48 performs best with highest overall accuracy, followed by rule barn (J Rip) and the K-NN classifier with Naïve Bayes being least accurate classifier.

Table 1. Algorithms identified through exploratory study

S.No	Algorithm	Research papers with same algorithm
1.	Decision Tree (J48)	[8], [9], [4],[1],[7],[2], [3],[15], [6]
2	Random Forest and Random Tree	[4],[14], [15]
3.	Naive Bayes	[8], [11],[1], [15], [6]
4.	Multilayer Perceptron	[2], [3], [14]
5.	SMO	[4], [2],[12],[14]

3 Research Methodology

A systematic study and investigation of existing material and sources to establish new body of knowledge or devise new knowledge is called research methodology.

3.1 Research Design

The research design adopted is exploratory and predicted due to the nature of the study**.** An exploratory design was needed to conduct literature survey and identify factors used by researchers in past. Further, in order to develop predictive model, the data set has to go through an experimental set up under which

different algorithms are applied on the data set (sample). Then the algorithm that gives the best result is selected for model building. Thus, experimental design follows the exploratory design in this study.

3.2 Samples and Sampling Technique

A professional course was to be selected constructing the data set., professional course has to be selected. We have selected Master of Computer application (MCA) course as population. The first year of the course builds the necessary foundation as the students are selected from heterogeneous. Thus, the effect of learning environment of the Institution will be reflected after one year of the joining of the institution. Thus, predicting the third semester result will help us understand the extent to which foundation has been built. The experiment was conducted to predict the third semester results of all students affiliated to Guru Gobind Singh Indraprastha University.

3.2.1 Data Collection

A structured questionnaire is constructed using Google Doc which has been administered to the MCA students from all the institutions affiliated to Guru Gobind Singh University over Internet. The sampling technique used fall under the category of convenience sampling. A data of 1545 students was collected.

Performance prediction includes attributes pertaining to academic integration (AI) (Table 2), social integration (SI) (Table 3), and emotional skills (ES) (Table 4). The emotional skills attributes were calculated based on the responses to a set of questions. For this standard tool of Emotional skill assessment Process (ESAP) was used.

Table 2 Attributes Pertaining to Social Integration

Attribute	Values
Gender	Male, Female
Father's Education (FE)	Secondary, Senior secondary, Grad, Postgrad
Mother's Education (ME)	Secondary, Senior secondary, Grad, Postgrad
Father's Occupation (FO)	Government job, Private job, Business, Others
Mother's Occupation (MO)	Government job, Private job, Business, Housewife
Family Income (FI)	Low income Group (LIG)(<2 lakh per annum)
	Middle income group (MIG) (2 to 4 lakh per annum) Other
Loan	Yes, No
Early Life (Where a student has spent first 15 years of his life	Metro ,City, Village

Table 3 Attributes Pertaining to Academic Integration

Attribute	Value
Medium of Instruction at school level	English, others
Percentage of Marks in Secondary, Senior secondary, Graduation, Firstsem, SecondSem, Third Sem, Fourthsem	Below average BLAVG (<60) Average AVG (60 to less than 70) Above Average ABVG (70 to less than 80) Excellent EXCL(>=80)
Type of Graduation degree (GRADDEGTYPE)	Regular, Distance
Graduation Stream (GRADDEGSTREAM)	Computer Science(CS) Non Computer Science(NCS)
Gap year after Graduation (GAPYEAR)	Yes, No
Hours spent on academic activities (ACADEMICHRS)	Insufficient(UNSUF)<2Hrs Sufficient(SUF) 2-4Hrs Optimal > 4 Hrs.
On campus placement (ONCAMPUSPLACE)	Yes, No
Relevant work experience (RELWORKEXP)	Yes. No
Project	Yes, No

Emotional Skill attributes are assessed through Emotional Skill Assessment Process (ESAP) tool developed by (Nelson and Low, 2003), consisting of psychometric questions to judge various parameters. Following attributes were considered

Table 4. Emotional Skill Attributes

Attributes	Values
1. Assertion (Ability to communicate effectively, honestly, clearly)	
2. Empathy (Ability to care for others)	D: Needs to develop the skill(Absent)
3. Decision making (Ability to take informed decisions)	S: Skill is present but need to strengthen(Moderate)
4. Leadership (Ability to influence others)	E: Skill is present and Enhanced(Enhanced)
5. Drive Strength (Ability to set a goal and strive for it)	
6. Time Management (Ability to manage time for best productive use)	
7. Self Esteem (Ability to regard himself or herself)	
8. Stress management (Ability to work under stress)	

4. RESULT AND ANALYSIS

After the collection of data, it needs to be cleaned and formatted to make it suitable for modelling. Data was captured inExcel, cleaned and saved as Comma Separated Variable (CSV). Next Waikato Environment for Knowledge Analysis (WEKA) , a popular software for data mining was used for analysis J48, which is the decision tree algorithm version of WEKA, Two ensemble method algorithms Random Tree and Random Forest and Bayesian Net, a form of Naive Bays algorithms were applied to the data set.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 739-745 June 2018-Dec 2018 www.csjournals.com

Table 5 Results of application of Classification Algorithms J48 and Random Forest Tree

Algorithms	J48			Random Forest Tree		
	TP Rate	Precision	Recall	TP Rate	Precision	Recall
ABVG	0.976	0.819	0.976	0.935	0.935	0.935
EXCL	0.915	0.826	0.915	1.000	1.000	1.000
AVG	0.980	0.824	0.980	0.929	0.903	0.929
BAVG	0.923	0.828	0.923	0.944	1.000	0.944
Weighted Average	0.943	0.835	0.943	0.944	0.944	0.944
Correctly Classified	94.6%			94.4%		
In Correctly Classified	5.4%			5.6%		
Time to build the Model (Seconds)	0.01			0.03		

Table 6 Results of application of Classification Algorithms Random Forest Tree and Bayes Net

Algorithms	Random Tree			Bayesian Network		
	TP Rate	Precision	Recall	TP Rate	Precision	Recall
ABVG	0.935	0.973	0.935	0.688	0.726	0.688
EXCL	1.00	1.000	1.00	0.786	0.786	0.786
AVG	0.923	0.904	0.923	0.929	0.843	0.929
BAVG	0.949	0.923	0.949	1.000	1.000	1.000
Weighted Average	0.945	0.950	0.945	0.795	0.794	0.795
Correctly Classified	94.5%			79.53%		
In Correctly Classified	5.5%			20.47%		
Time to build the Model(Seconds)	0.03			0.01		

The performance comparison shows that Bayesian Network algorithm has least accuracy while other algorithms have almost comparable accuracy J48 (94.6%), Random Forest (94.4%), Random Tree (94.6%) the highest being that of J48 algorithm with minimum building time . Further the model building time for J48 is lowest and as it generates a single decision tree it can be easily converted to rules which will be simple Model for the performance prediction.

5 CONCLUSIONS

In this paper decision tree algorithms J48, Random Forest Tree, Random Tree have been applied along with Bayes Net Algorithm to classify educational data for predicting thirdSemester result of MCA students. J48 Tree has been found to be most suitable for thisproblem. This is supported by [9].In future more, Algorithms like Support Vector Machine and Neural Network canalso be compared.

References

1. Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, *9*(4), 832-837.
2. Bharadwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance ,International Journal Advanced Computer Science and application ,2 (6), 234-240.
3. Cheewaprakobkit, P. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 13-15).
4. Cortez, P. and Silva,P. (2008) Using data mining to predict secondary school student performance, Proc. 5th Annual Future Business Technology Conference, Porto, pp. 5-12.
5. Guruler, H., Istanbullu, A., &Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, *55*(1), 247-254.
6. Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72.Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.
7. Kovačić, Z. J., & Green, J. S. (2010). Predictive working tool for early identification of 'at risk'students.
8. Nghe, N. T., Janecek, P., &Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In 2007 37th Annual Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports (pp. T2G-7). IEEE.
9. Ogor, E. N. (2007, September). Student academic performance monitoring and evaluation using data mining techniques. In Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007) (pp. 354-359). IEEE.
10. Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology, 9(1), 72-79.
11. Ramaswami,M. and Bhaskaran R.,(2010), A CHAID based performance prediction model in educational data mining, " International Journal of Computer Science Issues, 7(1) 10-18
12. S. Huang, (2011) Predictive Modeling and Analysis of Student Academic Performance in an Engineering Dynamics Course,
13. Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., &Wani, E. (2011, April). Prediction of student academic performance by an application of data mining techniques. In International Conference on Management and Artificial Intelligence IPEDR (Vol. 6, No. 1, pp. 110-114).
14. Sen, E. Ucar, and D. Delen, (2012.) Predicting and analyzing secondary education placement-test scores: A data mining approach, Expert Systems with Applications, Vol. 39, No. 10, pp. 9468-9476
15. Shah, N. S. (2012). Predicting Factors that AffectStudents'academic Performance by Using Data Mining Techniques. Pakistan business review, 13(4), 631-638.
16. Suljić, M. and Osmanbegović, E.,(2012). Data mining approach for predicting student performance. *Economic Review*, *10*(1), 3-12.