Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance

M. Sivasakthi
Asst. Professor of Computer Science and Applications
SRM University, City Campus, Vadapalani, Chennai 600 026, INDIA
Sivasakthi.info@gmail.com

***Abstract* - Data mining has been successfully implemented in the business world but, its use in higher education is still comparatively new. Predicting students' performance becomes more challenging due to the huge volume of data in educational databases. This paper focus on predicting introductory programming performance of first year bachelor students in Computer Application course by a predictive data mining model using classification based algorithms. The collected data contains the students' demographics, grade in introductory programming at college, and grade in introductory programming at test which contains 60 questions. Collected data was applied on various classification algorithms such as Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA. As a result, statistics are generated based on all classification algorithms and comparison of all five classifiers is also done in order to predict the accuracy and to find the best performing classification algorithm among all. In this paper, a knowledge flow model is also drawn for all five classifiers and also this paper showcases the importance of Prediction and Classification based data mining algorithms in the field of programming education and also presents some promising future lines. It could bring the benefits and impacts to students, educators and the academic institutions.**

Keywords- Educational Data Mining; Classification; Prediction; Introductory programming

I. INTRODUCTION

Data mining is one of the most popular research areas, due to its attention among researchers in recent years. Data mining techniques have been widely applied almost in all fields to ananyse the data for classifications, predictions, decision trees, fuzzy rules and so on. There has been increasing research interest in use of data mining techniques to investigate in the ground of education.

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods from educational environments to explore the unique type of data and those methods to better understand students in settings in which they learn [1]. Application of EDM methods has been in discovering or improving models of a domain knowledge structure and studying pedagogical support. There has been a witnessed rapid growth of EDM in order to improve the students learning process. Academic performance is crucial factor in building the students' future [2] [3]. Predicting and analysis of students' academic performance is an indispensible milestone in educational environment. In Computing education attempts have been made to predict success in programming courses [4]. The main objective of this paper is to apply the data mining algorithms such as multilayer perception, Navie Bayes, SMO, J48, REPTree to classify and predict the novice programmers in programming education.

II. BACKGROUND AND PRIOR WORK

The International working group in EDM established a yearly international conference that began in 2008 and the Journal of Educational Data Mining in 2009. An Ample of studies have been conducted on EDM in order to discover the effect of using it on students' performance [2]-[12]. A study to predict students' academic performance, the result discovered that students' university performance is depending on unit test, Assignment, attendance and graduation percentage [9]. A study on predicting students' performance in programming course revealed that the factors like students' mathematical background, programming aptitude, gender, high school mathematics grade and locality influences their programming performance [10].

TABLE I
SAMPLE OF DATA MINING TECHNIQUES USED FOR PREDICTION OF STUDENT PERFORMANCE

Year	Author	DM technique	Accuracy
2011	Saurabh Pal et. al	Naïve bayes	Not assigned
		Naïve bayes	82.4%

2011	Sembiring et al.	K-Means	93.7%
		DecisionTree	80.2%
2012	Edin Osmanbegovet.al	Naïve Bayes	76.48%
		Multilayer Perception	71.2%
		J48	73.98%
2014	Vaibhav P. Vasani et. al	Naïve Bayes	86.4%
		J48	95.9%
2015	Parneet Kaur et al	Multilayer Perception	75%
		Naïve Bayes	65.13%
		SMO	68.42%
		J48	69.73%
		REPTree	67.76%
2015	Amriah mohammed shahiri et al	Neural Network	98%
		Decision Tree	91%
		SVM	83%
		K-nearest Neighbour	83%
		Navie Bayes	76%
2016	Gurmeet Kaur & williamjit Singh	J48	61.53%
		Navie Bayes	63.59%

The comparative study which compares the results of classification with respect to different performance parameters and they observed that Decision tree (J48) gives better result than Naïve Bayesian algorithm in terms of accuracy in classifying the data [11].

III. PROPOSED METHODOLOGY

Survey cum experimental methodology has been adopted in this study. Through the extensive search of literature and discussion with experts, the numbers of factors which influence on the performance on the students are identified. Those influencing factor are categorized and considered as input variables. For our study data was collected for a subject/course 'Programming in C' which is a first year course for BCA (Bachelor of Computer Application) in affiliated colleges of university of madras. Then the collected data was filtered out using manual techniques then the data is transformed into a structured format required by WEKA tool. After that features and parameters selection is identified.

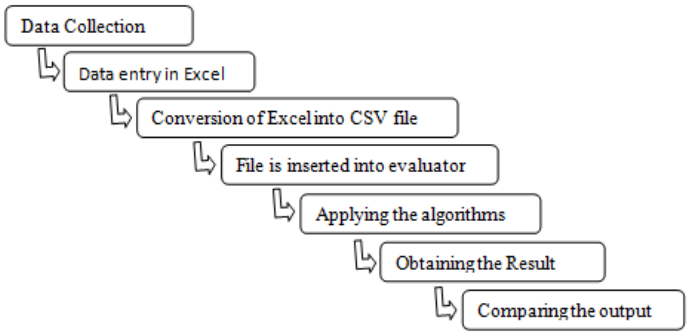

Fig. 1 Work methodology

A record of 300 students' data is used and student related variables are defined in the table-1 along with their domain values.

TABLE 2
STUDENTS RELATED VARIABLES

Sl. No	Variable Name	Description	Domain
1	Gender	Students' sex	{Male, Female}
2	HSC	Higher secondary studied	{CBSE, Metric, State Board}
3	Medium	Medium of Instruction	{Tamil, English}
4	Pvt-COH	Private coaching	{Yes, No}
5	School-Area	Area of School	{Urban, Semi-urban, Rural}
6	Grade-clg	Grade in C at college	{O,A+,A,B+,B,RA}
7	Grade-tst	Grade in C at Test	{O,A+,A,B+,B,RA}

TOOLS AND TECHNIQUES USED

In our study different data mining techniques are used to predict novice programmers. WEKA is used to apply the classification techniques and for predictions. The output is analysed with the following five classification algorithms.

J48: It is an implementation of C4.5 in WEKA. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced error pruning. J48 can use both discrete and continuous attributes, attributes with differencing lost and training data with missing attribute values.

REPTree: It uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

IV. RESULTS AND DISCUSSION

This section shows the results of various classifications shaped by WEKA and discuss the result with recent works in predicting students' performance. 10 folds cross validation test was supplied to those five algorithms with 'grade-tst' target attribute. The results have been obtained and observed the classifications' accuracy and matrices. The performance of those classifications have compared and shown in TABLE-III.

TABLE III
PERFORMANCE COMPARISON OF CLASSIFICATION ALGORITHM

	J48		REPTree		Multilayer Perception		SMO		Navie Bayes	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
B	0.905	0.704	0.889	0.741	0.911	0.759	0.905	0.704	0.600	1.000
A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.959	1.000
A+	1.000	1.000	0.973	1.000	1.000	1.000	1.000	1.000	1.000	1.000
B+	0.843	0.956	0.857	0.933	0.869	0.956	0.843	0.956	1.000	0.567
Weighted Average	0.923	0.920	0.921	0.920	0.934	0.932	0.923	0.920	0.902	0.845
Correctly classified instances	92.03%		91.03%		93.23%		90.03%		84.46%	
Incorrectly classified instances	7.97%		8.97%		6.77%		9.97%		15.54%	

Multilayer perception: It is a nonlinear classifier based on the Perception. A Multilayer Perception (MLP) is a back propagation neural network with one or more layers between input and output layer.

SMO (Sequential Minimal Optimization): SMO is a new algorithm for training Support Vector Machines (SVMs). Training a support vector machine requires the solution of a very large quadratic programming optimization (QP) problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically

Naïve Bayes: The Naïve Bayes is a simple probabilistic classifier. It is based on an supposition about mutual independency of attributes. The probabilities applied in the Naïve Bayes algorithm are calculated.

By looking at the TABLE-III, Multilayer Perception has highest prediction accuracy by 93% followed by J48 by 92%. Next REPTree by 91% and SMO by 90%. Lastly the method which is lower prediction accuracy is Navie Bayes by 84%. Comparison of performance accuracy is shown in below Fig.2.

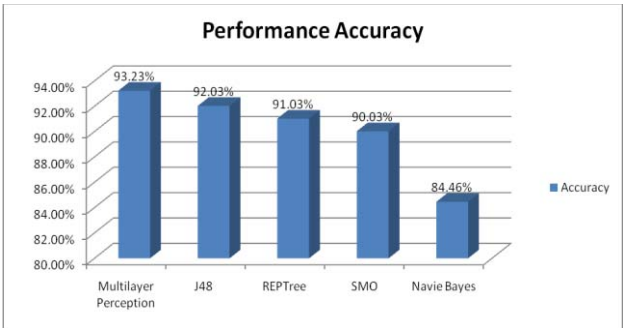

Fig.2 Comapsion of Performance accuracy

Multilayer Perception (MLP) has highest prediction accuracy by 93%. In MLP method, the factor that leads to the highest accuracy in predicting students' performance is students' grade. There is another study supporting this statement by which when they include grade prediction was about (71%) of performance accuracy [12].

Second higher prediction accuracy is J48 method b 92%. A study support this were considered attributes are SSC marks (percentages), mathematics (marks in mathematics), medium (language of instruction), and area (place of residence). However, the result decreased to (74%) performance accuracy when distance, scholarship, family were included as another feature [12].

Next is REPTree with the performance accuracy around 92%. In REPTree method, the factors that lead to the highest accuracy in predicting students' performance are medium of instruction, place of residence, private coaching and grade Point. There is a study which does not include those attributes; the result prediction was about 68% of performance accuracy [13].

Subsequently the SMO method is next prediction accuracy by 90%. A study which does not include are medium of instruction, place of residence, private coaching and grade Point attributes; the result prediction was about 68% of performance accuracy [13].

Last of all the Navie Bayes method is lower prediction accuracy by 84%. The variables used are CGPA, student demographic, high school background, scholarship, social network interaction.

Knowkedge flow model for those five classifications have been generated with help of the knowledge folw application of WEKA and which is shown in Fig.3. based on this model 'Model Performance chart has drwan and shown in Fig.4.

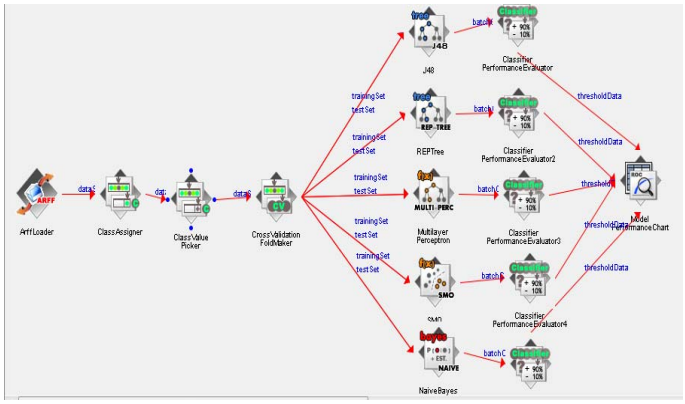

Fig.3 Knowledge flow model

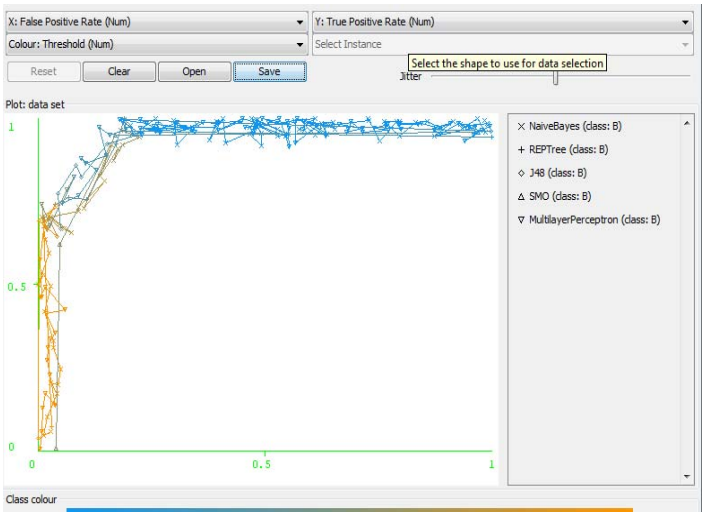

Fig.4 Model performance chart

V. CONCLUSION

In this paper, five supervised data mining algorithms were applied on the data set to predict programming performance of the students, were evaluated based on their predictive accuracy. The results indicate that the MLP performs best with 93% accuracy and therefore MLP proves to be potentially effective and efficient classifier algorithm. Also comparison of all five classifiers with the help of WEKA experimenter is also done, in this case also MLP proves to be best. Therefore, performance of MLP is relatively higher than other classifiers. A model performance chart is also plotted. This research may help the institutions to identify the students who are novice programmers in introductory programming, which further provide base for deciding special aid to them. However, it can be concluded that this methodology can be used to help students and teachers to improve student's introductory programming performance.

REFERENCES

[1] C. Romero, S. Ventura, "Educational data mining : a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, (Applications and reviews)*, Vol. 40, Issue 6, pp 601-618, 2010.

[2] R.S. Baker, A.T. Corbett,K.R. Koedinger, "Detecting Student Misuse of Intelligent Tutoring Systems". *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, pp 531-540, 2004.

[3] T. Tang, G. McCalla," Smart recommendation for an evolving e-learning system: architecture and experiment", *International Journal on E-Learning*, vol. 4, issue 1, pp 105-129, 2005.

[4] M. de Raadt, M. Hamilton, R.F. Lister, J. Tutty, B. Baker,I. Box, & M. Petre. "Approaches to learning in computer programming students and their effect on success". *Research and Development in Higher Education Series*, Vol. 28, pp 407-414, 2005.

[5] Saurabh Pal, "Data Mining: A Prediction For Performance Improvement Using Classification", *International Journal of Computer Science and Information Security*, Vol. 9, Isuue 4, pp 136-140, 2011.

[6] M. Sembiring, Zarlis , Dedy Hartama ,S. Ramliana , Elvi Wani, "Prediction Of Student Academic Performance By An Application Of Data Mining Techniques", *International Conference on Management and Artificial Intelligence*, Vol. 6, 2011.

[7] A. M. Shahiri et al , A Review on Predicting Student's Performance using Data Mining Techniques, *The Third Information Systems International Conference, Procedia Computer Science* 72, pp 414 – 422, 2015.

[8] G. Kaur, and S. Williamjit, "Prediction Of Student Performance Using Weka Tool." *An International Journal of Engineering Sciences, Vol. 17, 2016.*

[9] B. Suchita and K. Rajeswari, "Predicting Students Academic Performance Using Education Data Mining," *International Journal of Computer Science and Mobile Computing. IJCSMC*, Vol. 2, Issue. 7, pp 273 – 279, 2013.

[10] A. F. ElGamal, An Educational Data Mining Model for Predicting Student Performance in Programming Course, *International Journal of Computer Applications,* Vol. 70 Issue 17, 2013.

[11] P. Vaibhav, Vasani, D. Rajendra and Gawali, "Classification and performance evaluation using data mining algorithms*", International Journal of Innovative Research in Science, Engineering and Technology* , Vol. 3, Issue 3, 2014.

[12] Edin Osmanbegović, Mirza Suljić, "Data Mining Approach For Predicting Student Performance", *Economic Review – Journal of Economics and Business*, vol.10, Issue 1 2012.

[13] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in education sector." *Procedia Computer Science* 57,pp 500-508, 2015.

Sivasakthi M received the B.Sc and M.Sc in Computer Science from Periyar University, Salem, Tamil Nadu, India in 2004 and 2006 respectively, M. Phil in Computer Science from University of Madras, Chennai, in 2007, and he obtained Ph. D in Computer Science & Engineering Education from NITTTR (Ministry of HRD, Govt. of INDIA) affiliated to University of Madras in 2013.

He is working as Assistant Professor in SRM University, Vadapalani, Chennai, India since 2017. He has served as Assistant Professor in Loyola College, Chennai for couple of years. He is International Java Programmer Certified by Sun Micro System in 2008. His research is concerned with the pedagogical aspects of computer programming, data mining and cloud computing. He has contributed his research experience through 10 Journal publications, 12 Conference proceedings.