Mining Educational Data to Predict Learners' Performance Using Decision Tree Algorithm

Khor Ean Teng
Nanyang Technological University
Singapore

***Abstract*— Data mining is gaining increasing traction in the field of education as its applications in the education sector has increased over the past few years. Different data mining methods can be used to gain insights into educational data, including the uncovering of hidden patterns and prediction of output. The methods include classification analysis, association rule learning, anomaly or outlier detection, clustering analysis, and regression analysis. In this study, the classification analysis is used with decision tree algorithms to predict learners' performance. The findings reveal that the algorithm can be used to build a predictive model with good performance measure based on accuracy level, true positive (TP) rate, and false positive (FP) rate.**

Keywords—Educational Data Mining, Classification, Decision Tree Algorithm, Predictive Modeling, Learning Management System, Learners' Performance

I. Introduction

Educational data mining (EDM) is an emerging research field that is gaining the attention of education stakeholders because of its potential to enhance the teaching and learning process. EDM adapts and develops machine learning, data mining techniques, and statistical analyses to study educational data [1]. According to El-Halees [2], EDM used many methods like decision trees, support vector machine, naïve Bayes, neural networks, k-nearest neighbor and others to discover new knowledge.

The discovered knowledge of data mining is useful to various stakeholders of an education system. For example, learners are able to identify the learning task, resource and activities to enhance their learning while instructors are able to identify learner at risk, the most commonly made mistakes and to provide more feedback. Administrators, on the hand, are able to decide which courses to offer [3] and what new programme to launch.

New data-mining techniques embedded in learning management systems (LMS) extract information about the learning process from raw data [4]. According to Macfadyen and Dawson [5], learners' behavior data from LMS are recorded as background data. The data can be analyzed to gain insights on learners' learning progress.

There is a huge amount of data from web-based learning systems include LMS for analytic processing. According to García and Secades [6], the explicit data can be captured from LMS through any device for the activities performed by learners. The data will provide a clearer picture of the learning process to meet the needs of the learners.

In this study, the open educational dataset was used. The original source of the dataset is from authors in [7]. The dataset was collected from Kalboard360 LMS. Kalboard360 is a cloud-based LMS that has been designed to facilitate learning with leading-edge technology. The learners are provided a synchronous access to educational resources from any device with Internet connection [8].

II. Literature review

The aim of this research study is to generate a model to predict learners' performance based on demographical, academic background and behavioral attributes. Different data mining methods can be used to generate the predictive model [9]. In this research, the decision tree algorithm was used as it is one of the widely used classification techniques for prediction.

Decision tree algorithm was widely used due to its simplicity and comprehensibility to discover large or small data structure [10]. A set of IF-THEN rules can be converted and it is easily understood [11]. A decision tree which is in tree-shaped structures represent sets of a decision and the decision generates rules for the dataset classification [12].

A decision tree is a supervised classifier where it is generated from a training set. It is in the form of a tree structure and it contains data tuples. Each data tuple is represented by a class label and a set of attributes. The path from a root to a leaf can be followed based on the attribute values of the tuple and the leaf class is the predicted class of the particular tuple [3].

Kabra and Bichkar [3] suggested using a decision tree algorithm to build a model for predicting the performance of engineering learners. The model is based on their past performance data and it helps to identify the learners who are at risk or on the risk of failing so that warning can be given to improving their performance.

Ramaswami and Bhaskaran [13] examined the interrelation between variables with Chi-square Automatic Interaction Detector (CHAID) prediction model to predict the learners' performance at higher secondary school education. It is found that the medium of instruction, type of secondary education, academic performance of secondary education, living area and school location were seven important attributes to predict the outcome of learners' performance. The accuracy of the prediction model is 44.69%.

Merceron and Yacef [14] constructed the decision trees based on the data from web-based education system of Sydney University. If-then rules were generated to predict student marks he or she is likely to obtain. On the other hand, Kovacic [15] applied classification and regression tree (CART) and CHAID algorithms on student enrolment data to classify pass and fail students. The data was collected from students who studied information system (IS) at Open Polytechnic of New Zealand. The accuracy of the predictive model using CART and CHAID were 60.5% and 59.4 % respectively.

TABLE I.
THE ACCURACY OF THE DECISION TREE MODEL [10]

Result Accuracy	Attributes	Authors
91%	CGPA	Jishan, Rashu, Haque and Rahman [16]
90%	learners' extra-curricular activities, demographic, internal assessment	Elakia and Aarthi [17]
90%	learners' extra-curricular activities, demographic, CGPA, external assessment	Natek and Zwilling [18]
88%	psychometric factors, soft skills, extra-curricular activities	Mishra, Kumar and Gupta [19]
85%	external assessment	Bunkar, Singh, Pandya and Bunkar [20]
76%	internal assessments	Romero, Ventura, Espejo and Hervás [11]
73%	CGPA, learners' high school background, demographic, social network interaction, scholarship	Osmanbegović and Suljić [21]
66%	CGPA, internal assessment, extra-curricular activities	Mayilvaganan and Kalpanadevi [22]
65%	learners' demographic, high school background	Ramesh, Parkavi and Ramar [23]
65%	psychometric factors	Gray, McGuinness and Owende [24]

Table I illustrates the result accuracy of decision tree model with its attributes to predict the performance of learners. The related studies using decision tree include predicting the learners' performance at third-semester [19] and predicting the learners' suitable career based on their behavioral patterns [17]. Meanwhile, Mayilvaganan and Kalpanadevi [22] compared the classification models to predict the performance of learners while Gray, McGuinness and Owende [24] predict learners' progression in tertiary education by using a few classification models and the models are compared in terms of accuracy.

III. RESEARCH METHODOLOGY

The educational dataset was collected from kalboard360 cloud-based Learning Management system (LMS) using learner activity tracker tool which is known as Experience API (xAPI). xAPI is a new specification for learning technology [25] to track learning experiences. The tracked learning experiences are sent to a special xAPI-compliant database called a Learning Record Store (LRS). The LRS then reports what learners are doing.

Data were collected from 480 learners in two educational semesters: 245 learners in the first semester and 235 learners in the second semester. Out of 480 learners, 305 are male and 175 are female. For their education stages, 199 are lower level, 248 are middle school and 33 are high school. Majority of the learners (289 out of 481) absence of fewer than 7 days.

Pre-processing techniques were applied on datasets to remove the noisy data and feature selection was processed to reduce the number of attributes. Normalization mechanism was used whereby the numerical values were converted into nominal values for total marks of learners. Table II shows a class label that identifies learners' success into three categories based on learners' total mark: low-achieving learner (values between 0 and 69), middle-achieving learner (values between 70 and 89), high-achieving learner (values between 90 and 100). Data cleaning was performed by checking the missing value or irrelevant items of selected target data.

The dataset after pre-processing was exported to WEKA software. A .arff file called EduLMS was created and loaded into Weka Explorer. For this research, seven attributes (or predictor variables) is used to build the predictive model of learner performance. The attributes are categorized into three main groups: demographical category, academic background category and behavioral category (Table III).

TABLE II.
CLASS LABEL

Class Label	Description	Interval-Value
L	Low-Achieving Learner	0-69
M	Middle-Achieving Learner	70-89
H	High-Achieving Learner	90-100

TABLE III.

DOMAIN OF ATTRIBUTES

Category	Attribute	Data Type	Value
Demographical	Gender	Nominal	• Male • Female
Academic Background	StageID	Nominal	• Lower level • Middle school • High school
Behavioural	RaisedHands	Numeric	• 0 to 100
	VisitedResources	Numeric	• 0 to 100
	Announcements View	Numeric	• 0 to 100
	Discussion	Numeric	• 0 to 100
	AbsenceDays	Nominal	• Under-7 • Above-7

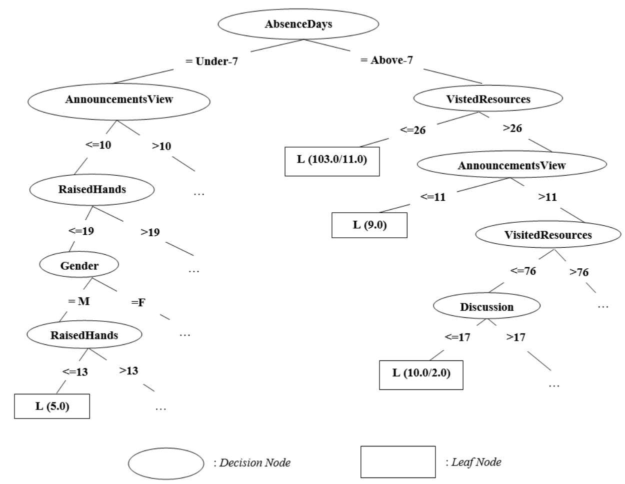

Fig. 1. Generated decision tree.

There are 16 decision tree algorithms in total. The algorithms include NBtree, ID3, Reptree, Simple CART and J48. Among them, J48 is known to be the best algorithm for the construction of a model [26]. Unlike the ID3 algorithm, the J48 algorithm does not need numeric attribute discretization. Discretization is a process to transform numeric attribute to a nominal value where the value of a numeric attribute is divided into a smaller number of intervals. Therefore, J48 was used in this study.

A 10-fold cross-validation was used to train and validate the model after which its performance was measured using a confusion matrix. A confusion matrix consists of information of actual and predicted classifications [27] and it illustrates the accuracy of the solution to the classification problem [28]. The generated model is based on the J48 algorithm and is in the form of a decision tree.

IV. ANALYSIS AND RESULTS

Fig. 1 shows parts of the decision tree constructed from EduLMS.arff file. The model achieves a 73.54% accuracy rate in classifying the instances correctly. In other words, 353 out of 480 instances are correctly classified. The feature like AbsenceDays is found to be the most significant indicators in predicting learners' performance.

The rules generated for low-achieving learner class from this tree are stated below:

1. IF AbsenceDays=Under-7 AND AnnoucementsView<=10 AND RaiseHands<=19 AND Gender=M AND RaisedHands <=13 THEN Class=L
2. IF AbsenceDays=Above-7 AND VistedResources<=26 THEN Class=L
3. IF AbsenceDays=Above-7 AND VistedResources>26 AND AnnouncementsView<=11 THEN Class=L
4. IF AbsenceDays=Above-7 AND VistedResources>26 AND AnnouncementsView>11 AND VisitedResources<=76 AND Discussion<=17 THEN Class=L

Table IV illustrates the results of the confusion matrix. Out of 127 low-achieving learners, 108 are classified as 'L'. Hence, the true positive (TP) rate and the false positive (FP) rate of class 'L' are found to be 0.805 and 0.076 respectively. For middle-achieving learners, 157 out of 211 are classified as 'M'. Therefore, the TP rate is 0.744 and the FP rate is 0.264. There are 88 out of 142 high-achieving learners are classified as 'H' with 0.620 TP rates and 0.086 FP rates. Table V shows the class-wise accuracy.

TABLE IV.

CONFUSION MATRIX OF 3 PREDICTED CLASS

		Predicted Class		
		H	M	L
Actual Class	H	88	53	1
	M	28	157	26
	L	1	18	108

TABLE V.

CLASS WISE ACCURACY OF 3 PREDICTED CLASS

Class	True Positive Rate	False Positive Rate
H	0.620	0.086
M	0.744	0.264
L	0.850	0.076

V. CONCLUSION

The study reports an overall correct classification rate of 73.54%. With a correct classification ratio of 73.54% achieved, this study concludes the potential use of the J48 decision tree algorithm to construct the predictive model to predict learners' performance. The constructed predictive model is based on learners' demographic (Gender), academic background (StageID) and behavioural (RaisedHands, VisitedResources, AnnouncementsView, Discussion, and AbsenceDay). Among the attributes, AbsenceDay was found to be the most significant predictors. The study involved 7 attributes and 480 instances. The accuracy of the model can be improved by adding more attributes and more instances. Based on the results of the confusion matrix, the TP rate and FP rate of the model are 0.850 and 0.076 respectively for the "L" class. The model is able to predict the learners who are likely to fall into low-achiever class. Those learners can be provided in-time intervention to improve the overall course success rate. Basically, there is no learner who absence more than 7 days scores 90% and above. In the future, ensemble methods like bagging and boosting will be studied to improve the modelling process and obtain better predictive performance. Besides the confusion matrix, precision and recall measures will be used to evaluate the classification of correctly classified instances and wrongly classified instances.

REFERENCES

[1] L.C. Liñán and Á.A.J. Pérez, "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution," International Journal of Educational Technology in Higher Education, vol. 12, pp. 98-112, 2015.

[2] A. El-Halees, "Mining students data to analyze e-Learning behavior: A Case Study," 2009.

[3] R. Kabra and R. Bichkar, "Performance prediction of engineering students using decision trees," International Journal of Computer Applications, vol. 36, pp. 8-12, 2011.

[4] M.Á. Conde, Á. Hernández-García, and A. Oliveira, "Endless horizons?: addressing current concerns about learning analytics," in Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality, 2015, ACM.

[5] L.P. Macfadyen and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," Computers & Education, vol. 54, pp. 588-599, 2010.

[6] O.A. García and V.A. Secades, "Big Data & Learning Analytics: A Potential Way to Optimize eLearning Technological Tools," in Proceedings of the International Association for Development of the Information Society, 2013, Prague, Czech Republic.

[7] E.A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in Proceedings of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015, IEEE.

[8] E.A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict Student's academic performance using ensemble methods," International Journal of Database Theory and Application, vol. 9, pp. 119-136, 2016.

[9] C. Kittidecha and K. Yamada, "Application of Kansei engineering and data mining in the Thai ceramic manufacturing," Journal of Industrial Engineering International, vol. 14, pp. 1-10, 2018.

[10] A.M. Shahiri, W. Husain, N.A. Rashid, and "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.

[11] C. Romero, S. Ventura, P.G. Espejo, and C. Hervás, "Data mining algorithms to classify students," in Proceedings of the 1st International Conference on Educational Data Mining, 2008, Québec, Canada.

[12] B.K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," International Journal of Advanced Computer Science and Applications, vol. 2, pp. 63-69, 2012.

[13] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining," IJCSI International Journal of Computer Science Issues, vol. 7, 2010.

[14] A. Merceron and K. Yacef, "Educational Data Mining: a Case Study," in Proceedings of the 12th International Conference on Artificial Intelligence in Education, 2005, Amsterdam, The Netherlands.

[15] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data," 2010.

[16] S.T. Jishan, R.I. Rashu, N. Haque, and R.M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," Decision Analytics, vol. 2, pp. 1, 2015.

[17] G. Elakia and N.J. Aarthi, "Application of data mining in educational database for predicting behavioural patterns of the students," International Journal of Computer Science and Information Technologies (IJCSIT), vol. 5, pp. 4649-4652, 2014.

[18] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," Expert systems with applications, vol. 41, pp. 6400-6407, 2014.

[19] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," in 4th International Conference on Advanced Computing & Communication Technologies (ACCT), 2014, IEEE.

[20] K. Bunkar, U.K. Singh, B. Pandya, and R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification," in 9th International Conference on Wireless and Optical Communications Networks (WOCN), 2012, IEEE.

[21] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," Economic Review, vol. 10, pp. 3-12, 2012.

[22] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," in Proceedings of the International Conference on Communication and Network Technologies (ICCNT), 2014, IEEE.

[23] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: a statistical and data mining approach," International Journal of Computer Applications, vol. 63, pp. 35-39, 2013.

[24] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in Proceedings of IEEE International Advance Computing Conference (IACC), 2014, IEEE.

[25] V.G.M. Ramirez, C.A. Collazos, F. Moreira, and C. González, "Relation between u-learning, connective learning, and standard xAPI: a systematic review," in Proceedings of the XVIII International Conference on Human Computer Interaction, 2017, ACM.

[26] M. Pandey and V.K. Sharma, "A decision tree algorithm pertaining to the student performance analysis and prediction," International Journal of Computer Applications, vol. 61, pp. 1-5, 2013.

[27] P. Singh and S. Verma, "Software Fault Prediction Model for Embedded Systems: A Novel finding," International Journal of Computer Science and Information Technologies, vol. 5, pp. 2348-2354, 2014.

[28] T.R. Patil and S. Sherekar, "Performance analysis of Naive Bayes and J48 classification algorithm for data classification," International Journal of Computer Science and Applications, vol. 6, pp. 256-261, 2013.