A Learning Analytics Approach to Model and Predict Learners' Success in Digital Learning

Khor Ean Teng
Nanyang Technological University
Singapore

Looi Chee Kit
National Institute of Education
Singapore

Learning analytics methods are widely applied in the educational field to gain insights on hidden patterns from educational data. Methods like predictive learning analytics are used to identify and measure patterns in learning data and extrapolate future behaviours. It can be used to enable the learners to be more self-aware of their learning behaviours and to enable the instructor to take appropriate actions informed by the trace of data. Thus such methods can empower learners as they progress through online training, and allows them to be self-regulated in order to solidify their learning and develop positive habits that will enhance their learning experiences. This paper reports on the use of a popular decision tree classification algorithm using behavioural features from a public domain dataset to develop a predictive model for predicting learning performance. Among the five behavioural features, we find that the measure of *visited resources* provides the most discriminating rules in the classifier.

Keywords: learning analytics, data mining classification, learners' success, learning behaviour, digital learning

1.0 Introduction

Learning analytics (LA) is an emerging field where it involves intricate analytic techniques to enhance teaching methods and learning activities (Bharara, Sabitha, & Bansal, 2018). With LA, education data can be converted into useful information and thereon to motivate actions to support learning experiences (Dyckhoff, Zielke, Bültmann, Chatti, & Schroeder, 2012). The analytics can gain hidden knowledge and insights about learners and to optimize learning (Romero & Ventura, 2010).

With the application of big data techniques in the educational sector, vital information such as learning behaviour can be extracted to understand the learners better. Learning behaviour features were included in the development of predictive models to predict learners' success and retention. For instance, Smith, Lange, and Huston (2012) included the number of times the learners log in to LMS and the number of times the learners engage in the material as features in predicting learners' performance in the course.

Learning behavioural features play a vital role in the explanation learners' successful (and unsuccessful) learning process. Recent research provides strong evidence that learner behaviour in online environments may predict learner success (Essa & Ayad, 2012). For example, Khor (2019) found that there was a significant relationship between learners' learning behaviour and academic performance. There is an improvement of the model accuracy when using learning behavioural features. This research work developed a learner's academic performance model based on learning behavioural features.

Various data mining techniques and algorithms have been implemented in predictive analytics to develop a model for predictions. For example, data mining techniques had been applied to forecast success in a course in Intelligent Tutoring Systems (Hämäläinen & Vinni, 2006); predict learner final marks based on Moodle usage data (Romero, Ventura, Espejo, & Hervás, 2008); and predict learner final grade based on features extracted from logged data (Minaei-Bidgoli, Kashy, Kortemeyer, & Punch, 2003). The techniques for predicting and analyzing student performance include decision tree methods like C4.5, RepTree and Cart, k-nearest neighbour, sequential minimal optimization, multi-layer neural networks and clustering methods. The advantage of a decision tree method is that the model provides readable rules that enable some intuitive understanding of the prediction mechanisms unlike other black box models. Hence, this study used the decision tree method to test whether it was an accurate predictor of learner' performance based on learning behaviour.

2.0 Methodology

In this study, a learner's academic performance model was developed based on learning behavioural features available from a public domain dataset. Table 1 illustrates the details of learning behavioural features. The popular decision tree classification algorithm was used to build the model. A decision tree is a classifier to classify an instance by following a path of satisfied conditions from the root until it reaches an end node (Romero et al., 2008). The algorithm develops methods to explore the unique types of educational data and the techniques are helpful to understand learners better.

Table 1: Descriptions and Data Types of Learning Behavioural Features

Features		Description	Data Type
Learning Behavioural	*raised_hand*	The number of times the learners asked questions.	Numeric
	visited_resources	The number of times the learners download the course materials	Numeric
	announcements_view	The number of times the learners view the announcement.	Numeric
	discussion	The number of times the learners participate in a discussion.	Numeric
	learner_absence	Learner's absence days.	Nominal

2.1 Description of Dataset

The source of the dataset was from Amrieh, Hamtini, and Aljarah (2016). The dataset was collected from Learning Management system (LMS) event log data using Experience API (xAPI). Data were collected from 500 learners in two educational Semesters and cleaned to remove 20 records with missing values from the data set. Of the 480 learners, 305 were male and 175 were female. For their education stages, 199 were lower level, 248 were middle school and 33 were high school.

2.2 Modelling Process

The modelling process for this study is presented in Figure 1. Data pre-processing techniques were performed on the dataset. Data pre-processing is the first step in the process of modelling to convert raw data into an appropriate form and comprises data cleaning, data transformation and feature selection (Miksovsky, Matousek, & Kouba, 2002).

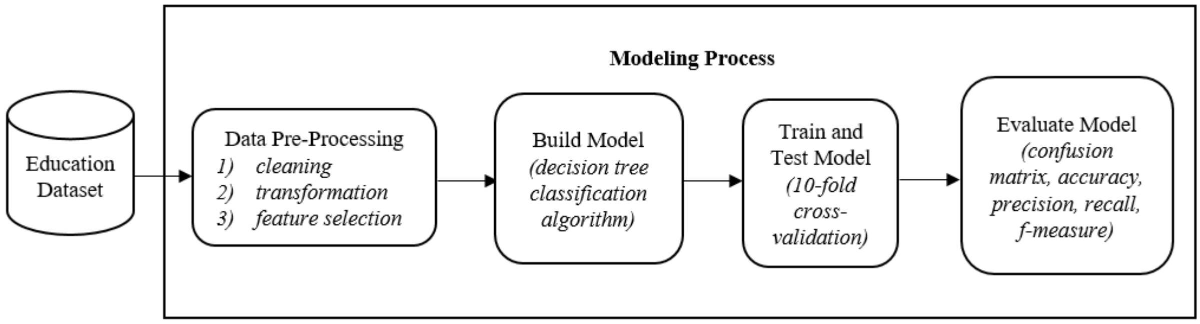

Figure 1: Summary of Modelling Process

Data cleaning was performed to check for the missing value of selected target data and to remove the noisy data. Data transformation was conducted to convert from a numerical (continuous) attribute to a nominal (categorical) value where the value of a numeric attribute is divided into a smaller number of intervals. In this study, learners' total final marks were converted from the numerical values to nominal values. Learners' success was categorised into three groups based on learners' marks: low-performer class (L) (marks between 0 and 69), moderate-performer class (M) (marks between 70 and 89) and high-performer class (marks between 90 and 100).

For feature selection, univariate feature selection was applied to increase the level of accuracy. Univariate feature selection chooses the best features based on univariate statistical tests. The score indicates the relationship of the features with the output variable. The higher the score, the stronger the relationship. Data mining classification process was carried out after the process of data pre-processing. A predictive model was then built using decision tree classifier. The inputs of the model included (1) *raised_hand*, (2) *visited_resources*, (3) *announcements_view*, (4) *discussion* and (5) *learner_absence*. The output of the models was *class_label* (academic performance).

Random sub-sampling 10-fold cross-validation was performed to categorise the dataset into training and testing groups. After running the model multiple times in randomized environment, an average result was produced (Smith et al., 2012). The performance of the developed model was then measured using a confusion matrix. A confusion matrix reports the classification of the actual and predicted class. The model was evaluated further with the value of accuracy, precision, recall and f-measure from the result of the confusion matrix.

3.0 Analysis and Results

The chi-squared scoring functions used in the univariate feature selection process is summarized in Table 2. These scores were used to identify the vital features of learning success in order to build the predictive model. As observed in Table 2, the feature *visited_resources* contained the highest weight, followed by *raised_hand*, *announcements_view*, *discussion* and *learner_absence*.

Table 2: Chi-squared Scoring Functions

Feature Selection	***raised_hand***	***visited_resources***	***announcements_view***	***discussion***	***learner_absence***
chi-squared (score function)	4124	4700	2618	809	135

Figure 2 shows parts of the constructed decision tree. It illustrates the rule-based classification generated for the high-performer class. The model achieved a 68.54% accuracy rate in classifying the instances correctly, meaning that 329 out of 480 instances were correctly classified. The results of the confusion matrix are presented in Table 3.

The performance of the developed model was also illustrated in terms of true positive (TP) rate, false positive (FP) rate, precision, recall, f-measure for the class label, L, M, and H respectively (Table 4). Out of 127 low-performers, 106 were classified as ‘L’. Hence, the TP rate and the FP rate of class ‘L’ were found to be 0.835 and 0.059 respectively. For middle-performers, 123 of 211 were classified as ‘M’. Therefore, the TP rate was 0.583 and the FP rate is 0.223. There were 100 out of 142 high-performers were classified as ‘H’ with 0.704 TP rates and 0.207 FP rates.

Table 3: Confusion Matrix Result

		Predicted Class		
		L	M	H
Actual Class	L	106	19	2
	M	20	123	68
	H	1	41	100

Table 4: Classification Report

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
L	0.835	0.059	0.835	0.835	0.835
M	0.583	0.223	0.672	0.583	0.624
H	0.704	0.207	0.588	0.704	0.641

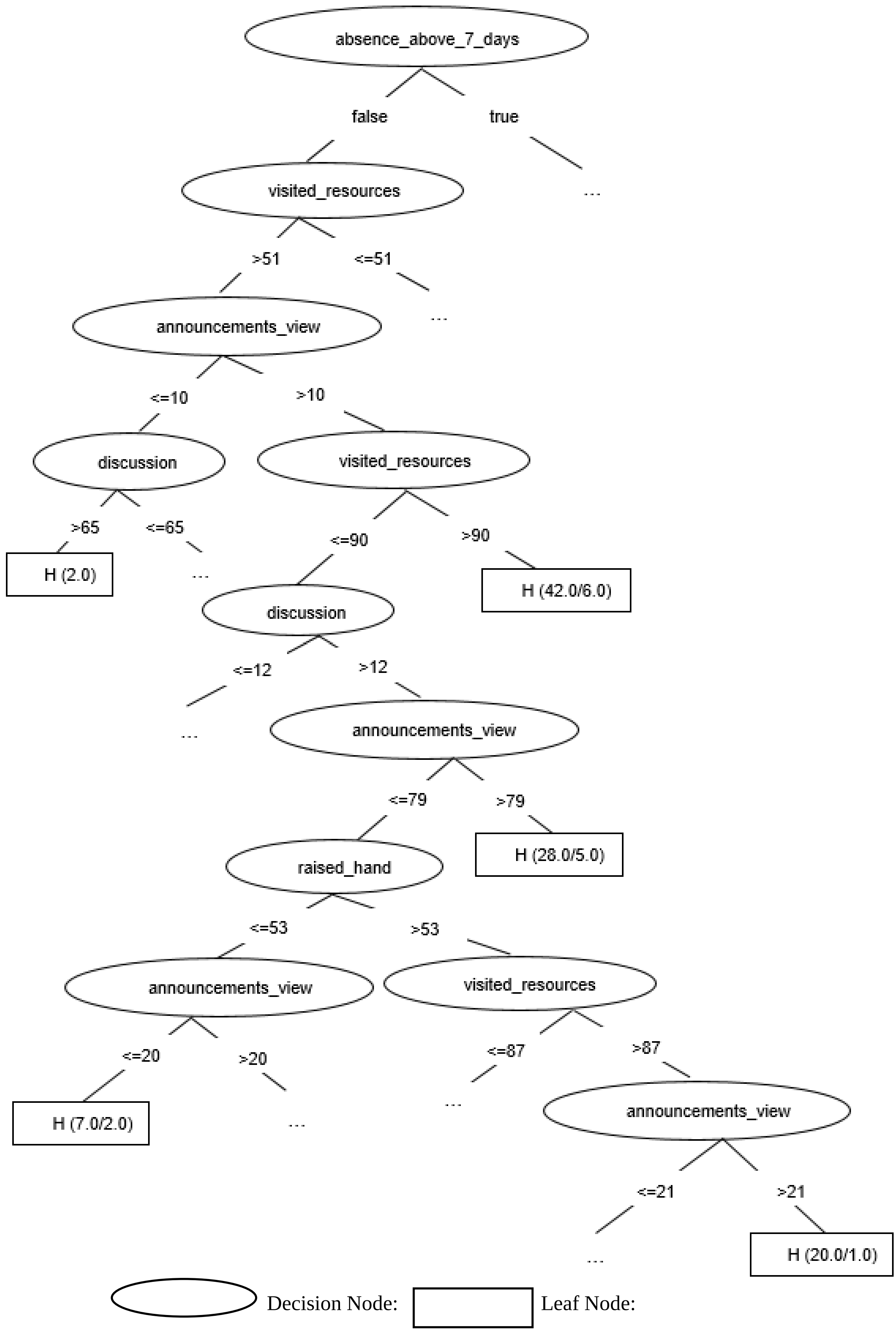

Figure 2: Constructed Decision Tree for High-Performer Class (H)

4.0 Conclusions

In this study, a predictive model was built using a decision tree classification algorithm. Specifically, C4.5 decision tree algorithms were applied to perform classifications on the dataset. For this dataset, the study found that the most important feature was *visited_resources*. The accuracy achieved using the decision tree classifier was encouraging, since the accuracy level is 68.54%. The class precision (0.835, 0.672, 0.588) and the class recall (0.835, 0.624, 0.641) for the three class labels (L, M, H) were appealing as well. The model can be used to make future predictions about learners' learning performance based on their learning behaviour. Proactive approaches and just-in-time interventions can be provided to at-risk learners to support their retention. While the field of learning analytics continues to develop, this study shows that classic classifier algorithms can still play their part in tackling specific classes of prediction problems in learning performance. Future works may include ensemble methods to gain better predictive process and enhance the performance of the model.

References

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). *Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance*. Paper presented at the Proceedings of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict Student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119-136.

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, *23*(2), 957-984.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Journal of Educational Technology & Society, 15*(3), 58-76.

Essa, A., & Ayad, H. (2012). Improving student success using predictive models and data visualisations. *Research in Learning Technology, 20*, 58-70.

Hämäläinen, W., & Vinni, M. (2006). *Comparison of machine learning methods for intelligent tutoring systems*. Paper presented at the International Conference on Intelligent Tutoring Systems.

Khor, E. T. (2019). *Predictive Models with Machine Learning Algorithms to Forecast Students' Performance*. Paper presented at the 13th International Technology, Education and Development Conference, Valencia, Spain.

Miksovsky, P., Matousek, K., & Kouba, Z. (2002). *Data pre-processing support for data mining*. Paper presented at the IEEE International Conference on Systems, Man and Cybernetics.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). *Predicting student performance: an application of data mining methods with an educational web-based system.* Paper presented at the FIE 2003 33rd annual Frontiers in education.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). *Data Mining Algorithms to Classify Students.* Paper presented at the Proceedings of the 1st International Conference on Educational Data Mining, Québec, Canada.

Sachin, R. B., & Vijay, M. S. (2012). *A survey and future vision of data mining in educational field.* Paper presented at the 2012 Second International Conference on Advanced Computing & Communication Technologies.

Smith, V. C., Lange, A., & Huston, D. R. (2012). Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. *Journal of Asynchronous Learning Networks, 16*(3), 51-61

Please cite as: Khor, E.T. & Looi, C.K. (2019 A Learning Analytics Approach to Model and Predict Learners' Success in Digital Learning. In Y. W. Chew, K. M. Chan, and A. Alphonso (Eds.), *Personalised Learning. Diverse Goals. One Heart. ASCILITE 2019 Singapore* (pp. 476-480).