Adoption of Feature Selection and Classification Techniques in a Decision Support System

Benilda Eleonor V. Comendador, Ariel M. Sison, and Ruji P. Medina

***Abstract*—This paper presents a decision support system prototype called eCourse Learning Analytics Decision Support System (eCLADSS) using J48 tree classifier and multiple linear regression models. The system identifies students who are falling behind in a course, notifies those at risk of not completing it, then informs the users the predicted grade a student is likely to obtain without intervention. The developed eCLADSS predicts the performance of the Learning Management System (LMS) users which may help the Distance Education (DE) students succeed in the blended learning approach being provided by the DE educators. It is a model-driven decision support system which provides a good platform for prediction model generation.**

***Index Terms*—Learning analytics, decision support system, classification techniques.**

I. Introduction

Research on learning analytics, knowledge discovery in databases and e-Learning technologies are gaining popularity in the academe because of their potential to improve services in the education domain. These innovations provide a ubiquitous tool and a powerful platform that may be used for higher educational institutions [1] and [2]. Learning analytics (LA) is the collection and analysis of usage data associated with student learning. Siemens defines Learning Analytics as the use of intelligent data, learner-produced data, and analysis models to discover information and social connections, and to predict and advise on learning [3]. The purpose of LA is to observe and understand learning behaviors for appropriate interventions. The reports generated by LA application can be very helpful for instructors (about student activities and progress), for students (feedback on their progress), and for administrators (e.g., aggregations of course and degree completion data) [4]. It can support teachers and students to take action based on the evaluation of educational data. However, the technology to deliver this potential is still in its infancy as well as the research on understanding the pedagogical usefulness of learning analytics [5]. LA synthesizes techniques from different related fields, such as Educational Data Mining (EDM), Academic Analytics, Social Network Analysis or Business Intelligence (BI). They all focus on tools and methods for exploring data coming from educational contexts. While Academic Analytics take a university-wide perspective, such as organizational and financial issues, Learning Analytics focus specifically on data about teaching and learning [6].

LA has five elements: 1) data which can be from a single source or a variety of sources; 2) analysis wherein its results are reported using a combination of visualizations, tables, charts, and other kinds of information display; 3) student learning which tell about students' learning performance;. 4) audience who utilized the information that LA returns; 5) interventions at the individual, course, department, or institutional level. The 2011 Horizon Report described the goal of learning analytics as enabling teachers and schools to tailor educational opportunities to each student's level of need and ability [7]. Unlike educational data mining, which emphasizes system generated and automated responses to students, learning analytics enables human tailoring of responses, such as through adapting instructional content, intervening with at risk students, and providing feedback.

Consequently, some educational institutions (e.g. California State University, Monash University in Australia) and several others were implementing Learning Management System (LMS) to manage the courses offered in the Internet. Based on the Moodle-LMS website there are 70,872 currently active sites that have registered from 234 countries. However, 44,397 of these have requested privacy and are not shown in their lists. Meanwhile, Picciano described that most LMS's provide constant monitoring of student activity whether there are responses, postings on a discussion board, accesses of reading material, completions of quizzes, or some other assessment [8].

Currently, little research has been conducted that focus on LMS learning analytics that will lead to the prediction of academic performance of the Distance Education (DE) learners. Thus, this paper focused on providing a decision support system prototype called eCourse Learning Analytics Decision Support System (eCLADSS) with the integration of Multiple Linear Regression (MLR) and J48 classification models of data mining.

II. Work Done / Contributions

Fig. 1 depicts the eCLADSS System Architecture that was used in this study.

It consists of four major phases such as 1) data pre-processing; 2) application of data mining techniques; 3) generation of student performance prediction model 4) then get appropriate decision for teaching and learning support.

Consequently, as illustrated in Fig. 2 the eCLADSS process model consisted of two stages: (1) the Data Mining

Manuscript received August 4, 2016; revised January 11, 2017. This work was supported in part by the Polytechnic University of the Philippines.

The authors are with Technological Institute of the Philippines, Graduate Programs, Quezon City, Philippines (e-mail: bevcomendador@pup.edu.ph, ariel.sison@eac.edu.ph, ruji.medina@tip.edu.ph).

doi: 10.18178/ijiet.2017.7.11.977

stage and (2) the Decision Support stage.

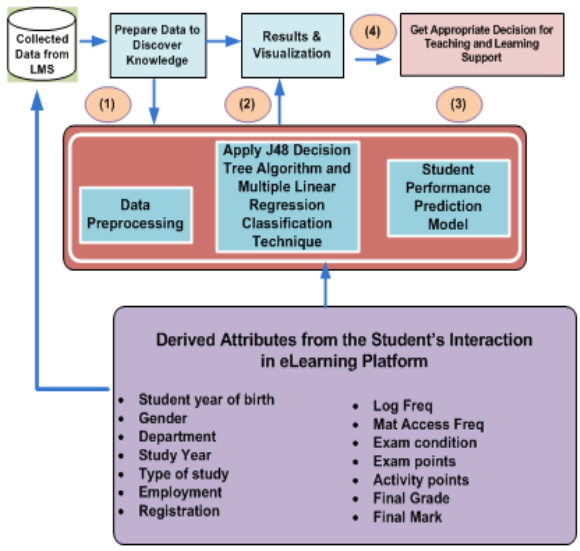

Fig. 1. eCLADSS system architecture.

The data mining stage presents the generation of the model in preparation for the decision support stage. It begins with (a) data preparation; (b) data selection and transformation; and (c) model generation. In this study, the data sets from the university Academic Institutions Management Systems (AIMS) and the students' history of accessing the Polytechnic University of the Philippines (PUP) eMabini Learning Portal-LMS were utilized. The LMS portal is powered by the Modular Object-Oriented Dynamic Learning Environment (Moodle) system. The user's log file and other student's related activities from the Moodle database consisting of 99,233 records were extracted**.** The authors used 248 records from the three (3) programs in the PUP Open University System.

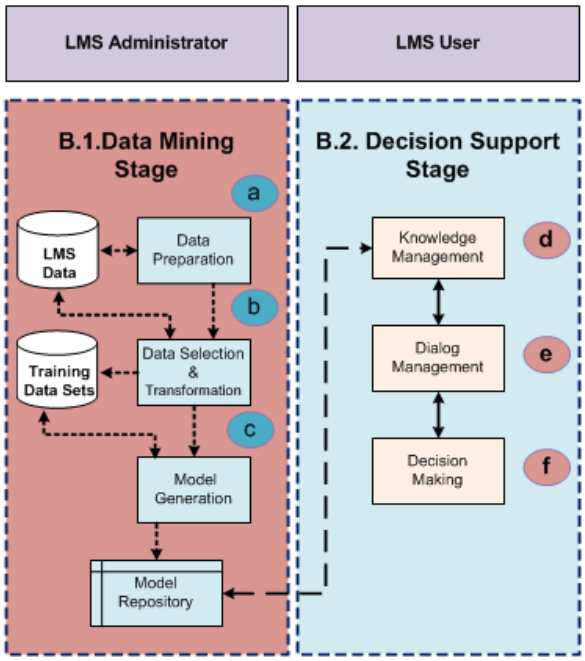

Fig. 2. Process model of eCLADSS.

On the other hand, the decision support stage begins with (d) knowledge management; (e) dialog management; and then (f) decision making. Knowledge management is the process where a model is selected and applied to the data. The knowledge is then presented to the LMS user in a manner that is easy to understand. However, dialog management is the process where the LMS user actually interacts with the system and inputs whatever data is needed to aid in the knowledge management. Results of the dialog will then facilitate in the decision making by the system user. Meanwhile, the LMS user is responsible for the model selection to enable him to analyze the data as well as the knowledge which would allow him to make appropriate decisions for teaching and learning support.

A. The Predictive Student Performance Model

Fig. 3 describes the process flow on how to generate the predictive student performance model. It consists of the following: 1) application of feature selection technique; 2) computation of attributes average and its rank; 3) application of J48 decision tree to classify the missing categorical value; and 4) application of the data mining regression algorithm to predict the student grade in numerical value.

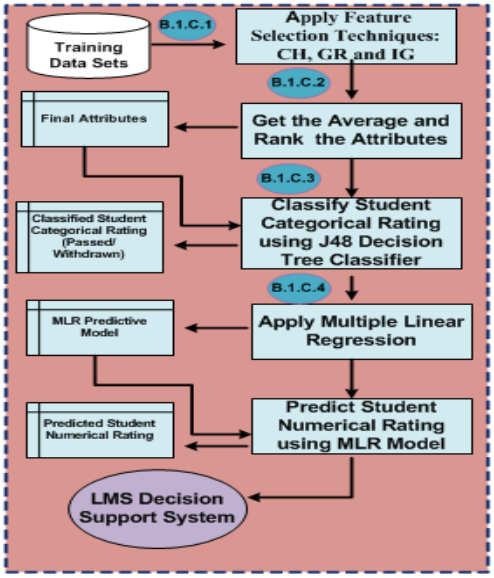

Fig. 3. Process flow of LMS predictive student performance model.

1) Application of feature selection techniques

Feature selection could be used as a pre-processor for predictive data mining to rank predictors according to the strength of their relationship with dependent or outcome variable [9]. In this work, the authors used the Chi-Squared Attribute Evaluation (CH), Gain-Ratio Attribute Evaluation (GR) and Information-Gain Attribute Evaluation (IG) in order to determine the value of a certain attribute in terms of interrelation among other attributes.

The average value of CH, IG and GR are taken in the final results of variable ranking to analyze the impact of each attribute on the construction of the student performance model.

As can be seen in Table I, the attribute activity_points garnered the highest average rate of 71.8. This result implies that attribute activity_points impacts output the most, and it showed the best performance in all of the three sets. The exam_points, exam _condition, study_year and log_freq attributes follow. However, the attribute registration got the lowest rank. This means an attribute getting a zero value has

no influence at all to the student performance model. As such, the top five attributes based on their average rank to generate the classifier model were used [10].

The authors carried out experiments in order to evaluate the performance and usefulness of the different classification algorithms for predicting students' final marks based on information in the students' usage data in the LMS and student profile. The discretization technique was applied during preprocessing tasks and tested the REPTree, CART and J48 algorithms on the data sets selected using 10-fold cross validation in WEKA. In this test, using J48 gain the highest accuracy rate of 97.17% [10].

TABLE I: AVERAGE AND RANK OF THE LMS USER'S ATTRIBUTES AFTER FEATURE SELECTION TEST

Attributes	Chi-Square (CH)	Info Gain (IG)	Gain Ratio (GR)	Average
Activity_Points	214.5	0.6	0.2	71.8
Exam_Points	155.1	0.4	0.2	51.9
Exam _Condition	113.4	0.3	0.6	38.1
Study_Year	21.2	0.1	0.1	7.1
Log_Freq	18.6	0.0	0.0	6.2
Mat_Access_Freq	16.4	0.0	0.0	5.5
Department	7.8	0.0	0.0	2.6
Age	4.3	0.0	0.0	1.4
Type_Study	3.8	0.0	0.0	1.3
Gender	1.2	0.0	0.0	0.4
Registration	0.1	0.0	0.0	0.0

2) Application of J48 decision tree classifier for categorical attribute

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. An important feature of a classification model is that it is built using part of the data, the training set, which is used to learn the model. In this subset all the attributes are known, including the class. After the model is built, it is used to assign a label to new records where the class attribute is unknown [11].

In this study, the algorithm that was used is J48 by WEKA, which is a version of earlier algorithm C4.5, wherein both continuous and discrete attributes can be used. This algorithm was developed by Ross Quinlan and it uses Gain ratio as an attribute selection measure. The J48 generates a decision tree using the concept of information entropy from a set of labeled training data. It examines the information gain to make decision. It is one of the decision trees which often used in classification and prediction because it is simple yet a powerful way of knowledge representation [12].

J48 works best in dealing with numeric attributes, missing data, noisy data, and generating rules from the tree. The algorithm works in heuristic based reasoning where the candidate cuts off a smallest number of instances on the numeric attributes. The J48 model used in eCLADSS was derived from our previous work that has an accuracy of 97.17% [10].

3) Application of multiple linear regression algorithm to predict numerical rating

Data mining algorithms are the mechanisms which create the data mining model and the main phase of the data mining process. It is an exclusive application of the classification rule. Regression involves smoothing data by fitting the data to a function. It could be linear or multiple, the linear involves finding the best line to fit two variables so that one could be used to predict the other, while the multiple one has to do with more than two variables [13].

In this study, the authors employed Multiple Linear Regression Algorithm to predict the Final_grade which involve three (3) numeric predictor variables such as activity points, exam_points and material access count. MLR is a data mining technique that allows the use of more than one input variables and allows for fitting of more complex models and was solved by the extension of least square method [14].

B. Functionalities of eCLADSS

The eCLADSS basically provides two major functionalities: the repeated generation of the Multiple Linear Regression (MLR) model and Performance Prediction for the LMS users to view and eventually make decisions as to what particular category of courses need special attention. The eCLADSS also used J48 decision tree classifier model for Learning Management System (LMS). It assists the learner to identify the most valuable influencer for learning outcomes. It determines how likely is a Distance Education (DE) learners to get a mark of "Passed" in a certain course which may offer vital information to the teachers and university administrators for program planning and learner support strategies. The eCLADSS Home Page is shown in Fig. 4.

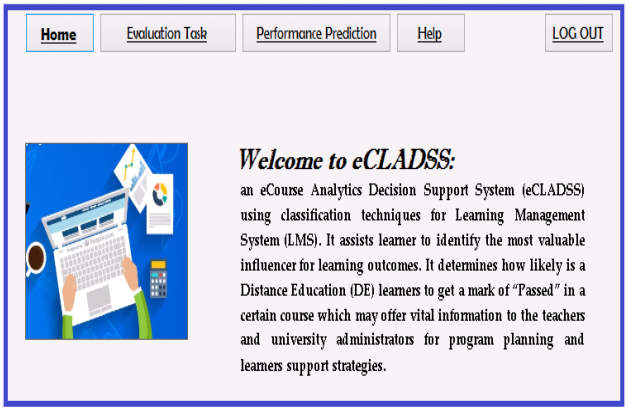

Fig. 4. eCLADSS home page.

The MLR model can be repeatedly generated by the LMS administrator using different training datasets which is selected from the data warehouse which is illustrated in Fig. 5. The selected dataset will serve as input to the generation of MLR model.

eCLADSS presents prediction by providing the predicted final rating of the DE Learners after having entered the necessary input as shown in Fig. 6 and Fig. 7.

The predicted computed mark is based on the multiple linear regression model set by the user from the model

repository. Only the LMS administrator has the authority to generate MLR models while the LMS user is granted an opportunity to select a model for the prediction of his LMS performance.

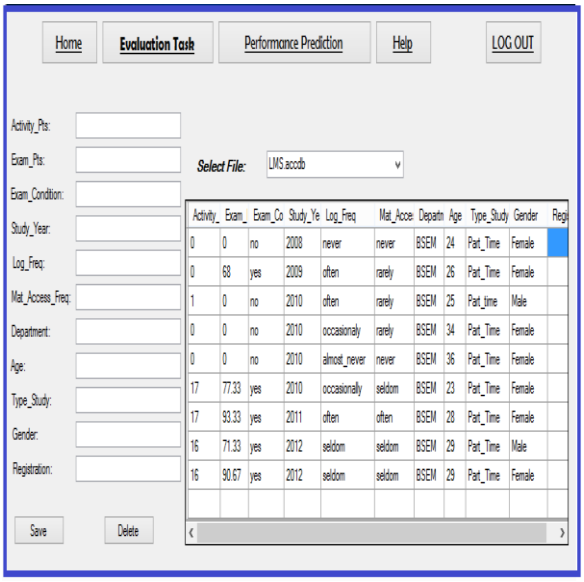

Fig. 5. Dataset selection interface for LMS administrator.

The authors used the top five attributes generated by the feature selection techniques to construct the predictive decision tree model. It consists of students' obtained score in the online activity, examination rating and its condition, year of admission, their frequency of log into the portal which served as valuable indicators to the predictive decision tree model [10]. To further aid in the prediction and support service of the system, J48 prediction model was used. An example of this functionality is shown in Fig. 6.

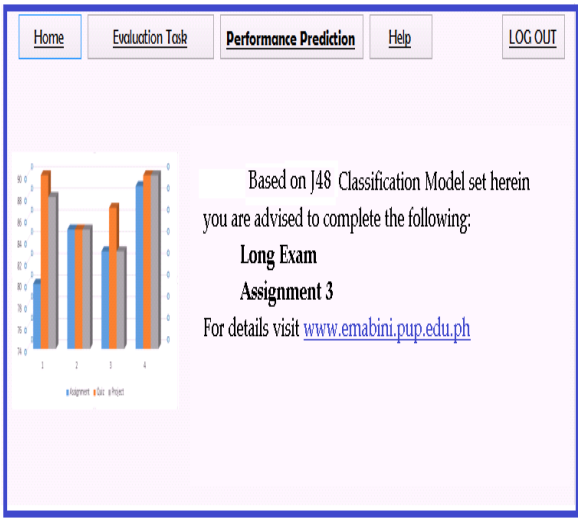

Fig. 6. Performance prediction interface with J48 classification model.

Thus, if we apply the generated multi-regression model in the LMS data sets, the student must get a ***score of 85 and 80*** in the Exam_pts and in the Activity_pts respectively, in order to obtain a Final_grade of ***85. 4617*** (see Fig. 7).

All models that have been selected and their corresponding predictions from the data entered by the users are stored in database. This database can be accessed by the LMS administrator as shown in Fig. 8 for the purpose of continued data analysis, model evaluation and model improvement.

Fig. 7. Performance prediction interface using MLR model for LMS user.

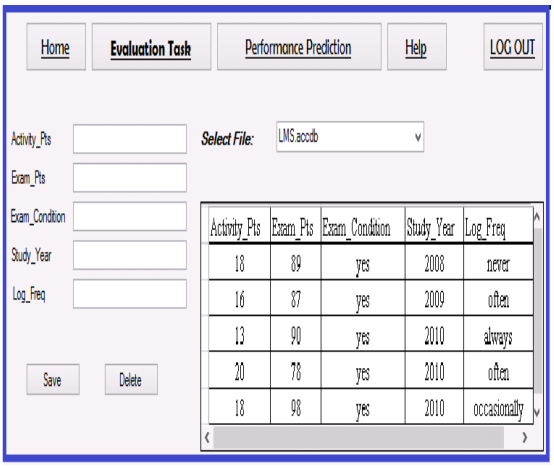

Fig. 8. Predictions from LMS user's data.

III. Conclusion

The output of the study is expected to offer the patterns or behavior of Distance Education (DE) students that may help them succeed in the blended learning approach being provided by the DE educators. The developed eCLADSS prototype using J48 and Multiple Linear Regression (MLR) models provide the LMS users to view and eventually make decisions as to what particular category of courses need special attention. The system assists the learner to identify the most valuable influencer for learning outcomes. eCLADSS also provides students notification to complete missed online activities and they can easily calculate predicted numerical rating.

eCLADSS was successful in integrating feature selection and classification techniques. It provides a good platform for generation of academic model that can be adapted by other educational institutions because of its model generation and selection procedures and user-oriented interface. In the future, the authors will extend their work by integrating to the system a feature that can provide a customizable attribute selection, enhanced interface that will help the user to have flexible interactive visual exploration of the accumulated data and to use other data mining techniques.

References

[1] V. Kumar and A. Chadha, "An empirical study of the applications of data mining techniques in higher education," *International Journal of Advanced Computer Science and Applications (IJACSA).* vol. 2 no. 3, pp. 80-84, March 3, 2011.

[2] D. G. Bayyou, "Cloud computing implementation in higher educational institutions using thin client," in *Proc. International Research Conference in Higher Education*, Manila, 2013, pp. 87-88, ISBN 978-971-781-037-9.
[3] G. Siemens. (2010). What are learning analytics? [Online]. Available: http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/
[4] M. Brown, "Learning analytics: The coming third wave," April 2011.
[5] L. Johnson, R. Smith, H. Willis, A. Levine, and K. Haywood. (2011). On the 2011 horizon report. Austin, Texas: *The New Media Consortium*. [Online]. Available: http://net.educause.edu/ir/library/pdf/HR2011.pdf
[6] J. P. Campbell and D. Oblinger. (2007). Academic analytics. [Online]. Available: http://connect.educause.edu/library/abstract/AcademicAnalytics/45275
[7] M. W. Johnson, M. J. Eagle, L. Joseph, and T. Barnes, "The EDM vis tool," in *Proc. the 3rd Conference on Educational Data Mining*, 2011, pp. 349–350.
[8] A. G. Picciano, "Big data and learning analytics in blended learning environments: benefits and concerns," *International Journal of Artificial Intelligence and Interactive Mulitmedia*, vol. 2, no. 7, 2014.
[9] Z. J. Kovačić, "Early prediction of student success: mining students enrolment data," in *Proc. Informing Science & IT Education Conference (InSITE)*, 2010.
[10] B. E. V. Comendador, L. W. Rabago, and B. T. Tanguilig III, "An educational model based on knowledge discovery in databases (KDD) to predict learner's behavior using classification techniques," in *Proc. IEEE International Conference on Signal Processing, Communications and Computing*, Hongkong, SAR, China, pp. 725-730, 978-1-5090-2708 August 7, 2016.
[11] C. E. Guar ń, "Data mining model to predict academic performance at the universidad nacional de Colombia," p. 20, 2013.
[12] P. Saini and A. K. Jain, "Prediction using classification technique for the students' enrollment process in higher educational institutions," *International Journal of Computer Applications*, vol. 84, no. 14, December 2013.
[13] Chapple. Regression. [Online]. Available: http://databases.about.com/od/datamining/g/regression.htm
[14] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques," p. 328.

Benilda Eleonor V. Comendador was a grantee of the Japanese Grant Aid for Human Resource Development Scholarship (JDS) from April 2008 to September 2010. She obtained her master of science in global information and telecommunication studies major in project research at Waseda University, Tokyo, Japan in 2010. She was commended for her exemplary performance in completing the said degree from JDS. She finished her master of science in information technology at Ateneo Information Technology Institute, Philippines in 2002. She is currently taking her doctor in information technology at the Technological Institute of the Philippines, Quezon city.

She is presently the program chair of the master of science in information technology of the Graduate School of the Polytechnic University of the Philippines (PUP) and designated as the chief of the Open University Learning Management System. She was the former chairperson of the Department of Information Technology of the College of Computer Management and Information Technology of PUP.

Her research interests include data mining, ubiquitous computing, eLearning and related technologies and data security. She presented several research papers in various international conferences and published them in the international journals.

Ariel M. Sison earned his doctor of information technology (DIT) at the Technological Institute of the Philippines Quezon City in 2013 and graduated with Highest Honors. He took up his master's degree in computer science at De La Salle University Manila in 2006 and obtained the BS computer science at Emilio Aguinaldo College Manila in 1994.

He is currently the dean, School of Computer Studies, Emilio Aguinaldo College Manila. His research interests include data mining and data security.

Dr. Sison is a member of International Association of Engineers (IAENG), Philippine Society of IT Educators and Computing Society of the Philippines. Currently, he is a Technical Committee Member of International Academy, Research, and Industry Association (IARIA) for International Conference on Systems (ICONS).

Ruji P. Medina is the dean of the graduate programs and concurrent chair of the environmental and sanitary engineering program of the Technological Institute of the Philippines in Quezon City. He holds a Ph.D. in environmental engineering from the University of the Philippines with sandwich program at the University of Houston, Texas where he worked on the synthesis of nanocomposite materials.

He counts among his expertise environmental modeling and mathematical modeling using multivariate analysis.