№ 2 (4), 2022 ISSN: 2682 – 9517 (print) ISSN: 2683 – 0930 (online)
INSTITUTE OF INFORMATICS AND INNOVATIVE TECHNOLOGIES

DATA MINING APPROACH FOR PREDICTING THE STUDENTS LEARNING OUTCOMES

Tran Thi Dung[1], Dzung Lai Manh[2], Cao Phuong Thao[3], Nguyen Le Minh[1]

[1]Division of Information Technology, Campus in Ho Chi Minh City, University of Transport and Communications, Ho Chi Minh City, Viet Nam.
dungtt_ph@utc.edu.vn; minhnl_ph@utc.edu.vn

[2]Faculty of Information Technology, University of Transport and Communications, Hanoi, Viet Nam. *dzunglm@utc.edu.vn*

[3]Faculty of Construction Management, University of Transport and Communications, Hanoi, Viet Nam. *thaocp@utc.edu.vn*

Abstract. ***Every year, many students are expelled because their test scores are below the allowable threshold of the university. To help universities and students know their learning status as well as improve their learning ability, this paper proposes an exam management system architecture and a method of predicting scores based on a regression model. Predictive method based on the relationship between subjects and student effort. The proposed method has been applied on the student learning data of the University of Transport and Communications, Vietnam and has improved the students' learning status.***

Keywords: ***score prediction, learning outcome prediction, regression model.***

I. INTRODUCTION

Predicting the learning outcome of students based on their responses to the status is a vital task for improve the student's education [1]. Rather than focusing on the approach to learning, the predicting learning outcomes sets the goals on what a student can accomplish [2]. Predictive Learning Analytics (PLA) is one of a research area which help lecturers improve course quality [3]. The learning outcomes allows for self-discovery and has the potential to move away from one-size-fits-all learning solutions [2]. However, these methods have been developed for and evaluated on the long courses, but these courses need frequency assessment and require the large number of enrolled students.

Recently, data mining can be applied in the field of education field, which helping to predict student learning outcomes, predict college admissions and possibly predict student learning outcomes. Superby et al [4] used questionnaires to collect the data including personal information, learning behaviors and perceptions of the students. Several machine learning methods have been applied on these data such as decision tree, random forest, neural network and linear discriminant analysis to analyze and predict the factors affecting the student's learning outcome. The data can be analyzed to find the factors that affect to the student learning outcomes when taking online learning courses [5 Ashby]. K-means algorithm has been applied to predict student learning behavior. The results obtained can help teachers adjust the lesson in the teaching process [6 Ayesha].

The data mining methods such as ID3, C4.5 and CART are also applied on data including student attendance, test scores, extracurricular activities to predict learning outcomes at the end of the semester [7, 8 Bharadwaj, Yadav]. Data mining also applied in education data to build personalized learning programs [9 Marie]. Lin [10] builds model to predict the students will have difficulty in learning so that they can provide timely support solutions. Dekker et al. [11] uses the Decision Tree method to build a model that predicts the proportion of students who may drop out after the first semester.

In this paper, we proposed a method of predicting scores based on linear regression model. The predictive method based on the relationship between subjects and

student effort, and the data collected from an exam management system. The experimental results on the data from course 54 to course 57 of the University of Transport and Communications, HCM City, Vietnam.

II. METHODOLOGY

Multivariate linear regression

Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis. Based on the number of independent variables, we try to predict the output. Multivariate regression tries to find out a formula that can explain how factors in variables respond simultaneously to changes in others. The equation for the multivariate regression model is as below.

$$Y = B_0 + B_1 * X_1 + B_2 * X_2 + .. + B_n * X_n \quad (1)$$

Where n represents the number of independent variables, β_0~ βn represents the coefficients, and x_1~x_n is the independent variable.

The cost function is a cost when the model differs from observed data. It can be calculated as the sum of the square of the difference between the predicted value and the actual value divided by twice the length of the dataset as equation 2. The smaller mean square error value means the better performance.

$$MSE = \frac{1}{2n}\sum(h_0(X)^{(i)} - Y^i)^2 \quad (2)$$

Minimizing the equation, the coefficients can be estimated as

$$\hat{B} = (X^T X)^{-1} X^T Y \quad (3)$$

Learning outcome prediction using multivariate linear regression

Consider an example with a student A, who has the course calculus 1 score of 8.5/10, and he want to know how much his calculus 2 score based on his status. Also, he wants to know the calculus 2 score if he tries his best. We can estimate this score using the collected data of calculus 1 and calculus 2 scores of the students from previous courses. From the collected data, the relationship between calculus 1 and calculus 2 will be calculated, then we can propose the regression function that can predict the calculus 2 score from calculus 1 score. The algorithm will be designed in two cases:

- Score prediction with current academic status, in this case, only previous scores data have used for training the regression model. To ensure the stability of the regression model, the training data is selected as a range of scores with small variation.
- Score prediction with the current academic with attemping behaviour status. For the linear model with tends to increase, which means the students have learning attemping behaviour, the training data is selected as the increasing or equal scores.

Table 1 shows several related subjects, based on this relation, the prediction result will more accuracy.

Table 1. Related subjects

Calculus 1	Calculus 2	
Linear Algebra	Probablity	
Data Structure and Algorithm	Introduction to Computing	Advanced Programming
Database Management System	Computer Ogranization and Architecture	
Object Oriented Programming	Introduction to Computing	Advanced Programming

In the table 1, the subjects in the row have relationship, the multivariate model have built based on this relationship.

Exam management system

We build the exam management system to store and collect data of the students. An overview of general architecture is shown in figure 1 The architecture components are broadly classified into 4 parts:

- The OAuth client (User Management) is used to control access to Practical Exam Management System (PEMS). The SSO (Single Sign On) authentication and authorization are implemented based on the Microsoft identity platform with the industry standard protocols OpenID Connect (OIDC) and OAuth 2.0. Students use the same Office 356 account to access system.

- The Exam scheduler allows academic staffs to manage practical exam schedules with basic functions such as creating manually or importing from schedule sheets extracted from training management system.

- The Data repository allows students to upload files of practical tests to the system, organize storage and manage them in an appropriated and secure way through several solutions: converting to compressed

format, normalizing file names, hiding the actual file paths.

- The Statistic & Report makes statistical reports on practical exam activities.

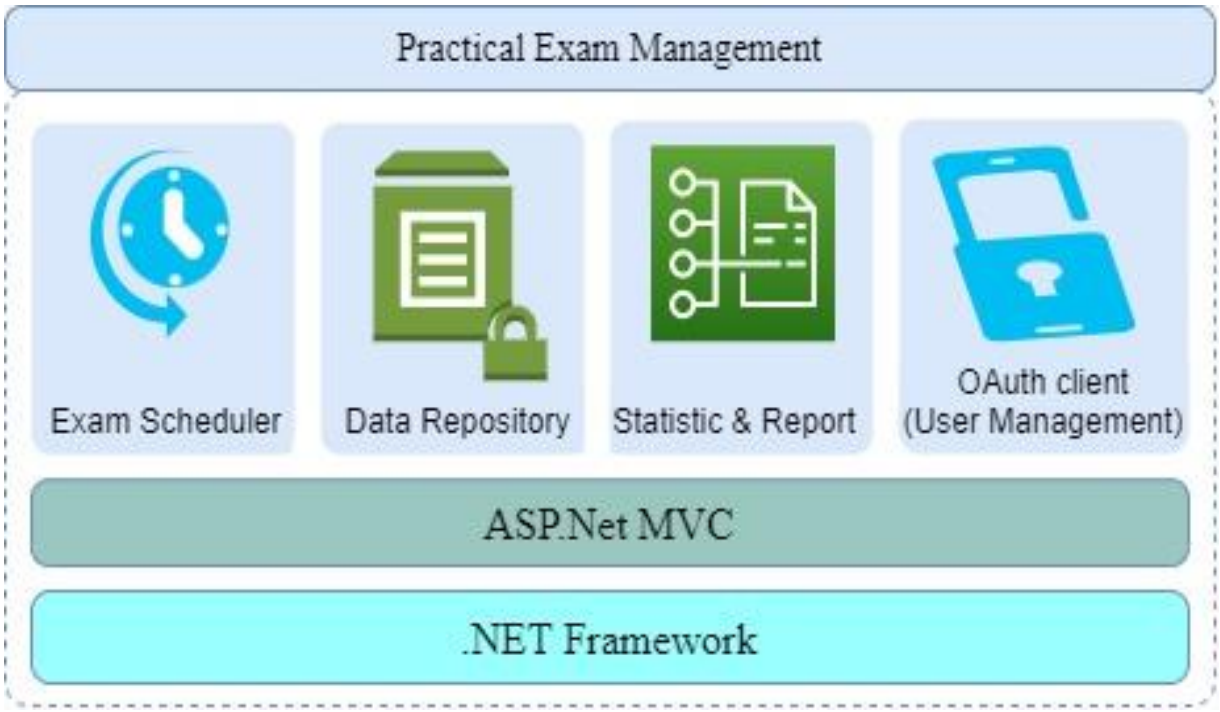

Fig. 1. General architecture of Practical Exam Management System

- The OAuth client (User Management) is used to control access to Practical Exam Management System (PEMS). The SSO (Single Sign On) authentication and authorization are implemented based on the Microsoft identity platform with the industry standard protocols OpenID Connect (OIDC) and OAuth 2.0. Students use the same Office 356 account to access system.

- The Exam scheduler allows academic staffs to manage practical exam schedules with basic functions such as creating manually or importing from schedule sheets extracted from training management system.

- The Data repository allows students to upload files of practical tests to the system, organize storage and manage them in an appropriated and secure way through several solutions: converting to compressed format, normalizing file names, hiding the actual file paths.

- The Statistic & Report makes statistical reports on practical exam activities.

III. EXPERIMENT RESULTS

The data used in this study is the scores data of all the subjects from students in four years of Information Technology, University of Transport and Communications, HCM City, Vietnam.

Table 2 shows the subject samples of the 2 semesters of the first year's student.

Table 2. Sample subjects' example

Subject	Score
Calculus 1	3.5
Introduction to Computing	4.7
English A1	9.5
English A2	5.4
Data Structure and Algorithm	5.4
Object Oriented Programming	8.5
Oracle	6.2
Data Mining	7.3
Artificial Intelligent	8

We run the prediction model in two cases, the first using linear regression model with current academic status. The second is prediction with the current

academic with attemping behaviour status. The prediction results of the two cases are shown in figure 2 and figure 3.

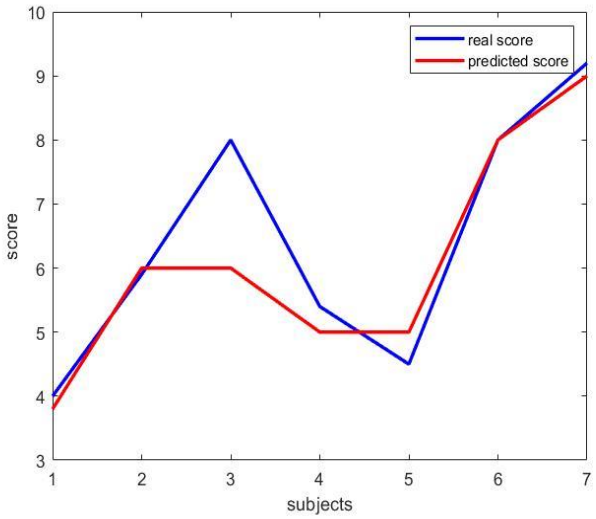

Fig. 2. Real and predicted score based on the student status

The result in the figure 2 shows that the prediction scores (the red line) do not improve much, even though lower than the previous semester. If the students do not attemping in their behaviour, they can not have the good scores.

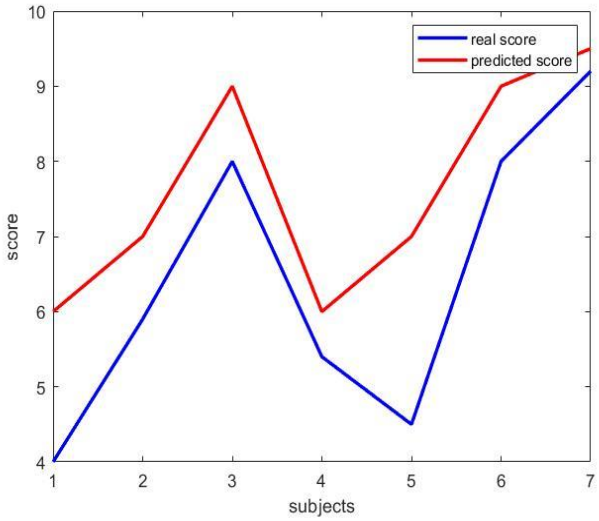

Fig. 3. Real and predicted score based on the student attemping

As in the figure 3, the prediction scores increase much compared to the figure 2. It means that if the students change the behavior positively, he can improve the scores. Also, the score in the common subjects is not change as much the scores in the major subjects.

In two cases, the experimental results show that the predicted scores are close to the real scores of the students, with an accuracy of up to 80%. Therefore, this model can be used by universities to have solutions to support students to achieve good results.

IV. CONCLUSIONS

This paper presents a method to predict student learning outcomes based on regression model. Although learning outcomes depend on many factors such as type of online or face-to-face learning, training programs. However, with the identification of related subjects and the addition of an attemping factor, the model has predicted the student learning outcomes with high accuracy. This method can help the univesity advise students to choose subjects and study methods to achieve high results.

ACKNOWLEDGEMENT

This research was supported by the project T2021-CN-002. We would like to gratefully acknowledge the support of University of Transport and Communication.

REFERENCES

[1]. Tianqi Wang, Fenglong Ma, Yaqing Wang, Tang Tang, Longfei Zhang and Jing Gao, Towards Learning Outcome Prediction via Modeling Question Explanations and Student Responses, Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pp. 693-701.

[2]. Chintan Donda, Sayan Dasgupta, Soma S Dhavala, Keyur Faldu, Aditi Avasthi, A Framework for Predicting, Interpreting, And Improving Learning Outcomes, https://arxiv.org/abs/2010.02629.

[3]. C. G. Brinton and M. Chiang, “Social Learning Networks: A Brief Survey,” in IEEE CISS, 2014, pp. 1–6.

[4]. Superby J. F., Vandamme J. P. and Meskens N., Determination of factors influencing the achievement of the first-year university students using data mining methods, Workshop on Education, 2006.

[5]. Ashby A., Monitoring Student Retention in the Open University: Detritions, measurement, interpretation and action, Open Learning, 19(1), pp. 65-78, 2004.

[6]. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar and M. Inayat Khan, "Data mining model for higher education system", European Journal of Scientific Research, Vol. 43, No. 1, pp. 24-29, 2010.

[7]. B. K. Bharadwaj and S. Pal., "Mining Educational Data to Analyze Student's Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69, 2011.

[8]. S. K. Yadav, B. K. Bharadwaj and S. Pal, Data Mining Applications: A Comparative Study for Predicting Student's Performance, International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.

[9]. Marie Bienkowski, Mingyu Feng and Barbara Means, Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics, Washington D. C.: U. S. Department of Education, 2012.

[10]. Lin S. H., "Data mining for student retention management", ACM Journal of Computing Sciences in Colleges, Vol. 27, No. 4, pp. 92-99, 2012.

[11]. Dekker, G., Pechenizkiy, M., and Vleeshouwers J. (2009), Predicting students drop out: A case study, In Proceedings of the 2nd International Conference on Educational Data Mining, pp. 41-50, 2009.