Predicting Student Performance Using Data Mining

Leo Willyanto Santoso[1], Yulia[2]
[1,2]*Informatics Department, Petra Christian University, Surabaya, Indonesia*
leow@petra.ac.id

***Abstract*— Supporting the goal of higher education to produce graduation who will be a professional leader is a crucial. Most of universities implement intelligent information system to support in achieving their vision and mission. One of the features of Intelligent Information System is student performance prediction. By implementing data mining method, this feature could accurately predict the student' grade for their enrolled subjects. Moreover, it can identify students that are at risk in failing a course and allow top educational management to take corrective actions. In this research, linear multi regression model was proposed to build model for every student. Based on the testing result on large set of students, courses, and activities shows that these models are capable of improving the performance prediction accuracy by over 15%.**

***Index Terms*— education; student; data mining; prediction**

I. INTRODUCTION

Education is a key to ending the poverty in developing countries. Education has power to change the people, communities, nation and human life. The government should pay more attention to the quality of education. Education is the responsibility of the stakeholders including government official, parent, and teacher. Education should be managed through national resources. Furthermore, higher education is important for social and economic impacts in society. The general mission of higher education institution is to produce student graduation who will be a professional leaders in their field and valuable for their communities and country. To achieve this mission, higher education institution should improve their quality of education. There are several factors affected the quality of education. The high level of student success and low failure rate students can reflect the quality of education. One of the major problems of higher education in the developing country, like Indonesia is the high rates of student drop out that has reached 10%. Another related problem is the long time that a student takes to complete their degree.

Nowadays, information technology is considered as important factor to improve the quality of education. This is the reason why many universities are investing a lot of budget to improve their academic information system [1].

Educational Data Mining (EDM) has emerged in the last decades due to the large volume of educational data that was made available. It is concerned with developing and applying data mining methods to detect patterns in large amounts of educational data, and to better understand students and their learning environments [2, 3]. Moreover, data mining and data warehousing technique have been increasingly implemented in the academic information system to analyze the vast amounts of student data [4]. Data mining is a tool to improve the quality of education by identifying the students who are at risk in their study. This information is very useful for top level management to take appropriate action for students who are considered to have a higher probability of failing academically or dropping out of university. The university could provide additional services and resources to the at-risk students. In addition, they need to develop innovative approaches to retain students, ensure that they graduate on a timely manner.

In this research, single regression model and multi regression model were implemented and investigated. This model could predict the students' grade by mining various course activities log (e.g., quizzes and assignments) in learning management system. An early warning system generates early warnings about struggling students who are most likely to failed a course or drop out of university. It is supposed to generate these warnings early enough in order to allow for intervention by offering suitable assistance for the students that are at risk. This system works by predicting a student's performance in the learning activities (e.g., assignments) within a course that they are enrolled in. They also predict the student's final grade in a course that they are enrolled in, or in courses that they will take in the next semester to fulfill their program requirements.

When students first enroll in a university, their university get the data about their performance in various high school subjects, test academic potential, and demographics. As the students proceed with their academic studies, more data are collected. The collected data like the student transcript and enrolled courses. The students can also access online learning management system (LMS), such as Moodle, Edmodo, Eliademi, ATutor or BlackBoard, at which they get access to the course materials. Through the LMS, students can also engage in forum discussions, contribute to the course content, engage in course activities such as online quizzes, and do other tasks. In this research, large dataset was extracted from the Petra Christian University's LMS. The name of Petra Christian University LMS is Lentera, based on Moodle. This dataset contains 486 courses, 7,563 students, and 109,231 activities.

The main contributions of this paper are as follows: (1) the designed system can cluster/segment the students into groups whose prediction models are relatively similar. By exploring these student' groups, knowledge on the factors that determine the students' performance are gained. 2) the proposed recommender system provides solution to improve the education quality using cutting edge technology.

The rest of the paper is organized as follows. Section 2 describes the literature review. Section 3 describes the multi-regression model that we used. Section 4 describes the dataset that we used along with the various features that we extracted. Section 5 provides the experimental evaluation and analysis of the results. Finally, Section 6 concludes this research.

II. Literature Review

Identifying at-risk students for taking appropriate actions can be addressed through evaluating collected students' academic performance data.

Decision tree technique was implemented to explain the interdependencies among the properties of drop out students [5]. This study also provides examples of how data mining technique can be used to improve the effectiveness and efficiency of the modeling process.

Dekker presented a data mining case study demonstrating the effectiveness of several classification techniques and the cost-sensitive learning approach. [6] In this system, cost-sensitive learning does help to bias classification errors towards preferring false positives to false negatives. Optimization should be done to improve the system.

Predictive analytic technique could be integrated with Learning Management System (LMS) to identify students who are in danger of failing the course in which they are currently enrolled [7]. Learning Analytic is considered can help teachers, educational managers, and students to predict course failure. Learning Analytic can help instructional material designers to better measure the quality of a course design and understand what works and what does not work. In addition, Learning Analytic can improve assessment of student performance by analyzing various indicators such as student postings and grades on assignments.

Data mining techniques for classifying students based on Moodle' usage data in a Learning Management System and the final marks obtained in the course was implemented [8]. The proposed system uses preprocessing tasks as discretization and rebalancing data. The author should consider how the quantity and quality of the data can affect the performance of the algorithms. Data with more information about the students, like student profile and curriculum should be incorporated.

Tensor factorization techniques for predicting student performance was proposed [9]. The author introduces a novel recommender system which can be used not only for recommending objects like tasks/exercises to the students but also for predicting student performance. The prediction results could be improved by applying more sophisticated methods to deal with the cold-start problems and building ensemble methods on different models generated from matrix and tensor factorization.

Several factors influencing the achievement of the first-year university students was determined [10]. The developed system can classify students into three groups: 'low-risk' students, with a high probability of succeeding; 'medium-risk' students, who may succeed; and 'high-risk' students, who have a high probability of dropping out. However, the combination of different prediction methods have not been addressed. This combination may lead to the improvement of the overall result.

With large volumes of student data, including enrollment, academic and disciplinary records, higher education institution could build big data and analytics system. Big Data can provide top level management the predictive tools they need to improve learning output for individual students as well ways ensuring academic programmes are of high-quality standards [11]. By designing programmes that collect data at every step of the students learning processes, universities can address student needs with customized modules, assignments, feedback and learning trees in the curriculum that will promote better and richer learning.

In this research, we investigate the linear multi-regression models for predicting the students' performance at various course activities in LMS.

III. Design and Implementation

In this part, the proposed model for prediction student performance will be discussed. This model uses multi-regression model [12]. Multi-regression is an extension of simple linear regression. As a predictive analysis, the multi-regression is used to explain the relationship between dependent variable and two or more independent variables. In this model, the grade $g_{s,a}$ for student s in activity a is formulated as.

$$\begin{aligned} g_{sa} &= b_s + b_c + \boldsymbol{p}_s^t W f_{sa} \\ &= b_s + b_c + \sum_{d=1}^{t} (p_{s,d} \sum_{k=1}^{n_f} f_{sa,k} w_{d,k}) \end{aligned} \quad (1)$$

Where:
b_s = student bias terms
b_c = course bias terms
f_{sa} = vector that holds the input features
l = number of linear regression models
W = matrix that holds the coefficients of linear regression
p_s = vector that holds the memberships of student s
$w_{d,k}$ = weight of feature k under the d^{th} regression model
$p_{s,d}$ = membership of student s in the d^{th} regression model

We compare the performance of a multi-regression model against the performance of a linear regression model. The linear regression model estimates the student grades as

$$g_{sa} = w_0 + \sum_{k=1}^{n_f} w_k f_k \quad (2)$$

where f_k is the value of feature k and the w_k's are the regression coefficients of the linear regression model.

In Fig. 1 can be seen the flow diagram of application design process. The initial stage is collecting data, then selecting data. Selection of data is needed, if there is missing value data, the data will be discarded. After doing data cleansing, then the data is divided into two namely the training data and test data with the percentage of each 70% for training data and 30% for the test data. The training data consists of prerequisite value as a predictor variable and predetermined value as a response variable. Test data just as the training data contains some prerequisite and predetermined value.

We used a dataset extracted from the Petra Christian University' Moodle. The main page of Petra Christian University' Moodle can be seen in Fig. 2. The dataset spans four semesters and it contains 486 courses, 7,563 students, and 109,231 activities. The courses belong to 21 different departments; each student has registered in around 4 courses. In this research, the activities refer to the assignments and quizzes

in Lentera.

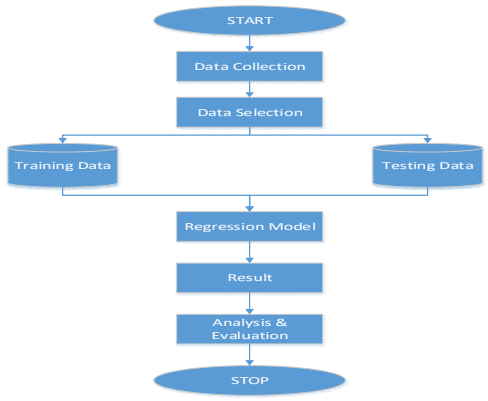

Figure 1: The Flow diagram of application design

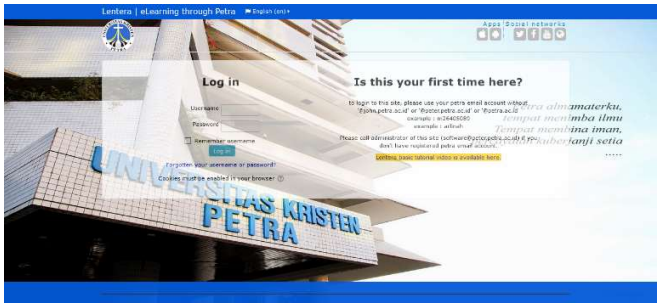

Figure 2: The main page of Lentera

For each student-activity pair (s,a), feature vector f_{sa} is constructed. There are three categories: student-centered features, activity-centered features and Lentera interaction features.

Student-centered features are features related to the student. There are two categories:

- GPA_total: The number of grade points a student earned in a given period of time.
- Grade_total: The average grade achieved over all of the pervious activities in the course.

Activity-centered features are features that relate to the activity of student in the Lentera LMS. Fig 3 describes the list of activities in Lentera. There are three categories:

- Activity_type: activity of student in order to interact with other student or teacher in Lentera, This can either be quiz or assignment.
- Course_level: The difficulty level of course. The range of value is 1, 2, 3 and 4. Value 1 means the difficulty of course is very low.
- Department: The department who offer the course.

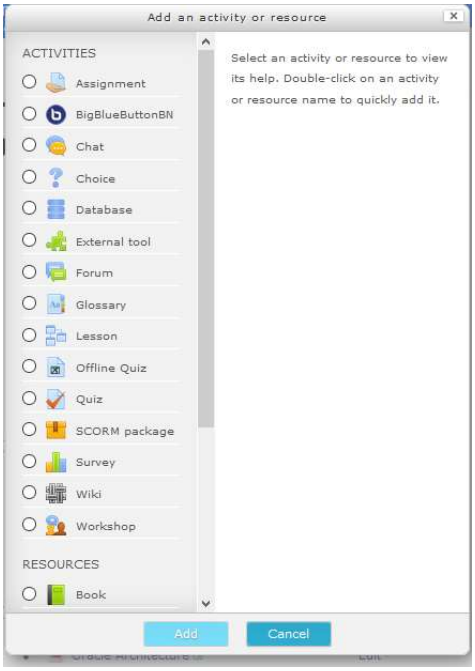

Figure 3: Activities in Lentera

Lentera-centered features describe the student's interaction with Lentera prior to the due date of the quizzes and assignments. These features were extracted from Lentera's log files and are the following:

- discuss_total: the number of discussion that posted by student.
- log_total: frequency of the student login to the Lentera
- time_total: total amount of time spent between login and logout
- read_total: the number of discussions' forum that are read by the student.
- viewed_total: the number of times the student viewed related material.

The dataset was divided into two subsets, namely training and test subsets containing 70% and 30% of the dataset respectively. The model was trained on the training set and evaluated on the test set. This process was repeated 5 times and the obtained results on the test set were averaged and reported. The model is evaluated in terms of the root mean squared error (RMSE) between the actual and predicted grades on the test set.

IV. Results and Discussions

The results and analysis are presented in this part. Moreover, the performance comparison between multi-regression model and single regression are discussed.

Figure 4: Statistics in Lentera

Fig. 4 shows the statistics in Lentera. It shows the number of active courses, students and activities in Lentera. The correlation between activities in Lentera (interaction between students with the Lentera features) and the predicted grades is discussed. To get the better result, the multi-regression models and the baseline model were trained 2 times.

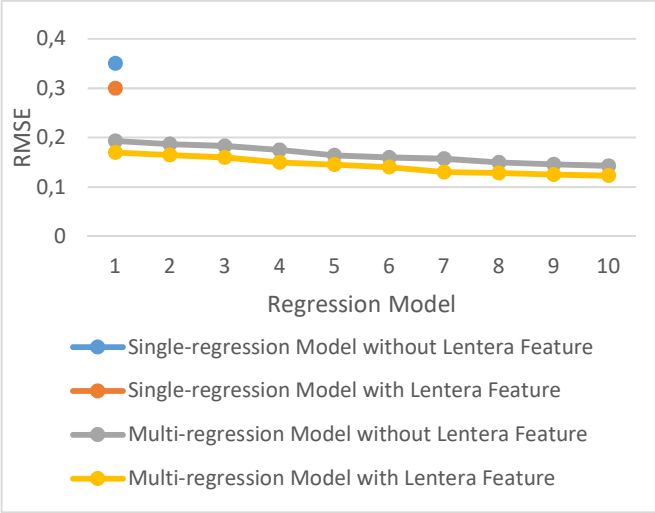

Figure 5: The graphic of regression model vs RMSE

Fig. 5 shows the graphic of the single regression and the multi-regression models with and without using Lentera-interaction features. It can be seen from this figure, the value of RMSE was change along this experiments.

It is clear from Fig. 5 that the RMSE of multi-regression model with Lentera features with one linear model is 0.17. On the other hand, the RMSE of single regression model is 0.3. By accompanying student-bias term and course-bias term, multi-regression model could better capture student performances in their course.

Fig. 5 illustrates that there is a decrement of RMSE obtained by the multi-regression model with increasing number of linear models. Using ten regression models, the obtained RMSE falls to 0.12.

Comparing the performance of the two multi-regression models in Fig. 5, we can see that the model that uses the Lentera features performs better than the one that does not use them. A multi-regression model with ten linear models gives and an RMSE of 0.143 without using the Lentera features and gives an RMSE of 0.12 using the Lentera features. The use of Lentera features lead to more drop in RMSE with increasing number of regression models. We believe this is because the model that uses the Lentera features have more student Lentera interaction information to learn from as the number of regression models increase.

V. Conclusion

In this research, multi-regression model to predict student performance in course was implemented. According to the testing result, multi-regression model performs better than single linear regression. Moreover, by increasing the number of linear regression model, the RMSE tends to decrease gradually. Finally, Lentera interaction features could improve the accuracy of prediction of student performance.

Acknowledgment

This research was supported by The Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Research Grant Scheme (No: 002/SP2H/LT/K7/KM/2017).

References

[1] L.W. Santoso and Yulia, "Analysis of the impact of information technology investments - a survey of Indonesian universities," ARPN JEAS. vol. 9, no. 12, pp. 2404-2410, Dec, 2014.
[2] www.educationaldatamining.org.
[3] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," Trans. Sys. Man Cyber Part C, vol. 40 no. 6, pp. 601–618, Nov, 2010.
[4] L.W. Santoso and Yulia, "Data warehouse with big Data technology for higher education," Procedia Computer Science, vol. 124, no. 1, pp. 93-99, 2017.
[5] M.N. Quadri and N. V Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," Glob. J. Comput. Sci. Technol., vol. 10, no. 2, pp. 2–5, 2010.
[6] G. W. Dekker, M. Pechenizkiy and J. M. Vleeshouwers, "Prediction student drop out: a case study," in *Proc.* 2nd International Conference On Educational Data Mining, Cordoba, Spain, 2009, pp. 41-50.
[7] R. Barber and M. Sharkey, "Course correction: using analytics to predict course success," in *Proc.* 2nd International Conference on Learning Analytics and Knowledge, Vancouver, Canada, 2012. pp 259-262.
[8] C. Romero, S. Ventura, P. G. Espejo and C. Hervás, "Data mining algorithms to classify students," in *Proc.* 1st Int. Conf. on Educational Data Mining, Montreal, Canada, 2008, pp 187-191.
[9] Ng. Thai-Nghe, L. Drumond, T. Horvath, A. Krohn-Grimberghe, A. Nanopoulos, and L. Schmidt-Thieme, "Factorization techniques for predicting student performance, in Educational recommender systems and technologies: practices and challenges, O. C. Santos and J.G. Boticario, Ed. IGI Global, 2011, pp. 129-153.
[10] J. F. Superby, J. P. Vandamme, and N. Meskens, "Determination of factors influencing the achievement of the first-year university students using data mining methods," in *Proc.* of 8th international conference on intelligent tutoring systems, Jhongli, Taiwan, 2006, pp. 37-44.
[11] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," British Journal of Educational Technology, vol 46 no. 5, pp. 904–920, 2015.
[12] A. Elbadrawy, R. S. Studham and G. Karypis, "Personalized multi-regression models for predicting students' performance in course activities," in *Proc.* 5th International Conference on Learning Analytics and Knowledge, Poughkeepsie, US, 2015, pp. 103-107.