TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 3, June 2020, pp. 1354~1360
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i3.14802

Predicting student performance in higher education using multi-regression models

Leo Willyanto Santoso, Yulia
Informatics Department, Petra Christian University Surabaya, Indonesia

Article Info

Article history:

Received Jul 23, 2019
Revised Jan 21, 2020
Accepted Feb 24, 2020

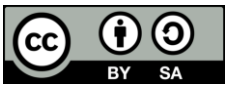
Keywords:

Data mining
Education
Multi-regression
Prediction
Student

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 (IIS) to support in achieving their vision and mission. One of the features of IIS is student performance prediction. By implementing data mining model in IIS, this feature could precisely predict the student' grade for their enrolled subjects. Moreover, it can recognize at-risk students and allow top educational management to take educative interventions in order to succeed academically. In this research, multi-regression model was proposed to build model for every student. In our model, learning management system (LMS) activity logs were computed. Based on the testing result on big students datasets, courses, and activities indicates that these models could improve the accuracy of prediction model by over 15%.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Leo Willyanto Santoso,
Informatics Department,
Petra Christian University Surabaya,
121-131 Siwalankerto St., Surabaya, East Java 60236, Indonesia.
Email: leow@petra.ac.id

1. 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 [2, 3]. It is concerned with developing and applying data mining algorithms to identify patterns in large amounts of educational data, and to better understand students and their learning environments [4-7]. Moreover, data mining and data warehousing technique have been increasingly implemented in the academic information system to analyze the vast amounts of student data [8, 9].

Data mining is a tool to improve the quality of education by identifying the students who are at risk in their study [10-12]. 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 [13, 14]. In addition, they need to develop innovative approaches to retain students, ensure that they graduate on a timely manner [15].

Some techniques have been developed to address this issue. However, these approaches ignore the different features of how students work together with the material/LMS' provided information, which could possibly be used to increase overall accurateness of prediction. In this research, single regression model and multi regression model were implemented and investigated. This model could predict the students' grade by mining different course activities log (e.g., tests 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 [16]. This dataset contains 486 courses, 7,563 students, and 109,231 activities.

The important 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 2 describes the multi-regression model that we used. Section 2 describes the dataset that we used along with the various features that we extracted. Section 3 provides the investigational evaluation and analysis of the results. Finally, section 4 concludes this research.

2. RESEARCH METHOD

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 properties interdependencies of drop out students [17]. This study also offers an example of how data mining technique can be used to increase the effectiveness and efficiency of the modeling processes. Dekker explained a data-mining case study demonstrating the usefulness of several classification methods and the cost-sensitive learning approach [18]. 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 [19]. Learning analytic is considered can support students, lecturers and educational managers to predict course failure [20]. Learning analytic be able to support instructional material designers to better measure the quality of a course design and understand what works and what does not work [21, 22]. Moreover, learning analytic can increase evaluation of student performance by investigating various indicators such as student activities and grades on assignments.

Data mining techniques for categorizing university students based on Moodle' usage data in a learning management system and the final marks achieved in the course was implemented [23]. The proposed system uses preprocessing tasks as discretization and rebalancing data. The author should consider how the data quantity and data quality can impact the performance of the algorithms. Information with more evidence about the students, like student profile and set of courses should be incorporated.

Tensor factorization techniques for predicting student performance was proposed [24]. 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 effecting the accomplishment of the freshman students was determined [25]. 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 [26]. Big data can provide top level management the needed analytical tools to improve learning output for individual students as well ways guaranteeing academic programmes are of high-quality standards [27]. By designing applications that gather data at every phase of the students learning processes, universities can address student needs with customized modules, feedback, and assignments in the syllabus that will stimulate better and richer learning. In this research, we investigate the linear multi-regression models to forecast the students' performance at various course activities in LMS.

2.1. Design

In this part, the proposed model for prediction student performance will be discussed. This model uses multi-regression model [28, 29]. 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 activities *a* is formulated as (1).

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

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

The performance comparison between a multi-regression model across a linear regression model was presented. The approximation of university student grade using linear regression model as;

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

where f_k is the rate of k and the w_k's are the regression coefficients.

In Figure 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 Figure 2. The dataset spans four semesters and it contains 486 courses, 7,563 students, and 109,231 activities. The courses belong to 21 different schools; each university student has registered in around 5 courses. In this research, the activities refer to the assignments and quizzes in 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 grades points a student earned in a given period of time.
- Grade_total: The average grade accomplished over the entirely of the past exercise in the course.

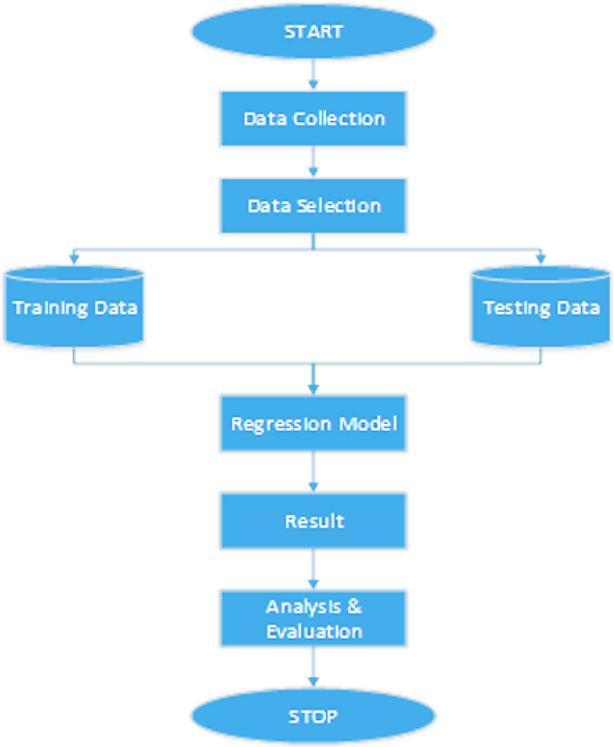


Figure 1. The Flow diagram of application design

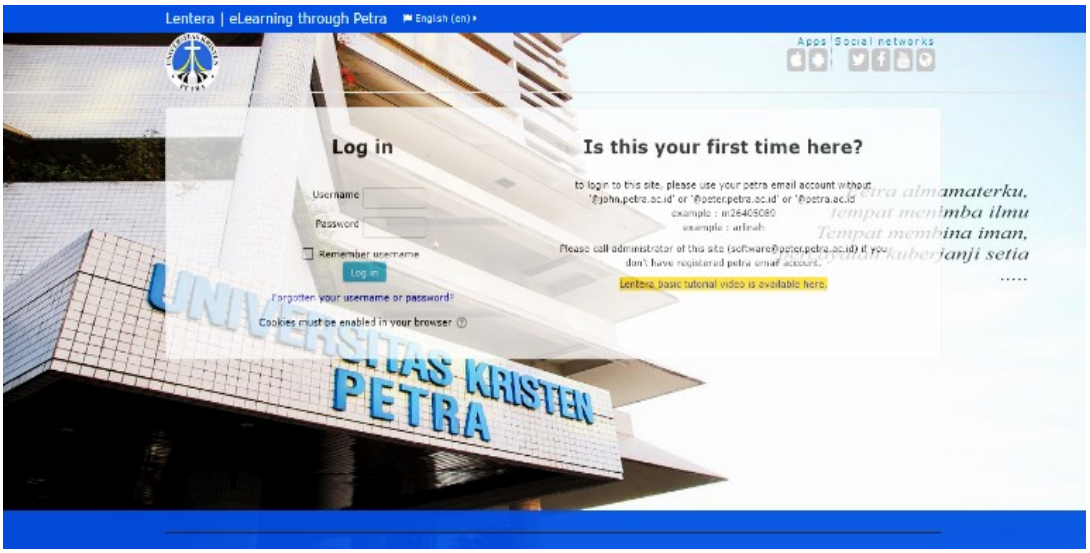


Figure 2. The main page of Lentera

Activity-centered features are features that relate to the activity of student in the Lentera LMS. Figure 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.

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 delivered by the student.
- viewed_total: the number of times the student viewed related material.

The dataset was divided into two subsets, namely training and testing subsets comprising 70% and 30% of the dataset respectively. The proposed model was trained on the training datasets and then evaluated on the testing datasets. This evaluation process was reiterated 5 times and the acquired results on the test datasets were calculated. The root mean squared error (RMSE) was used to assess the proposed model. It measures the difference between the actual and predicted grades on the test datasets.

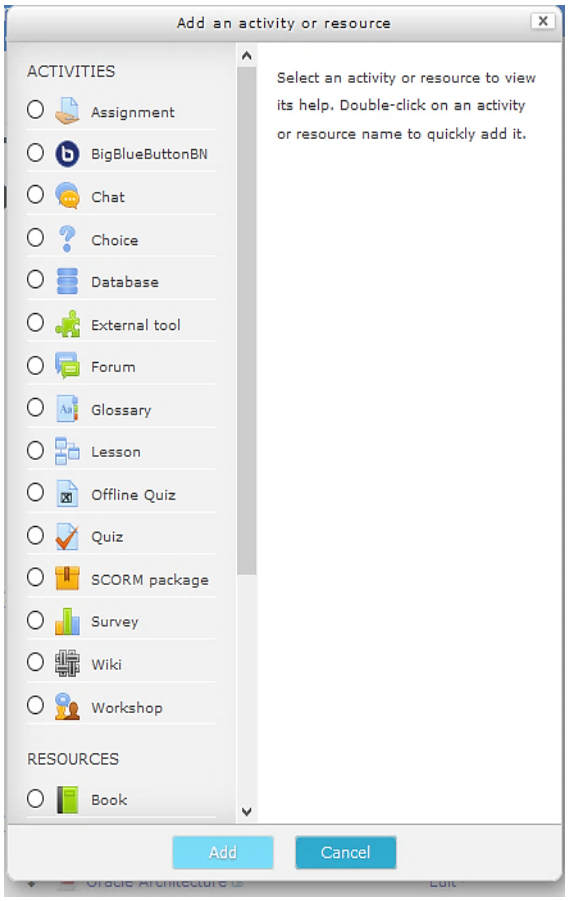


Figure 3. Activities in Lentera

3. RESEARCH FINDINGS AND ANALYSIS

This section presents the research findings and analysis. Moreover, the performance comparison between multi-regression model and single regression are discussed. Figure 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 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 Figure 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.

Figure 5 illustrates that there is a decrement of obtained RMSE by the multi regression model with increasing number of linear models. Using twelve proposed regression models, the acquired RMSE drops to 0.12. Comparing the performance of the two multi-regression models in Figure 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. From the evaluation, it can be concluded that it is because the proposed model that practices the Lentera features have extra student Lentera collaboration information to study from as the number of regression models increase.

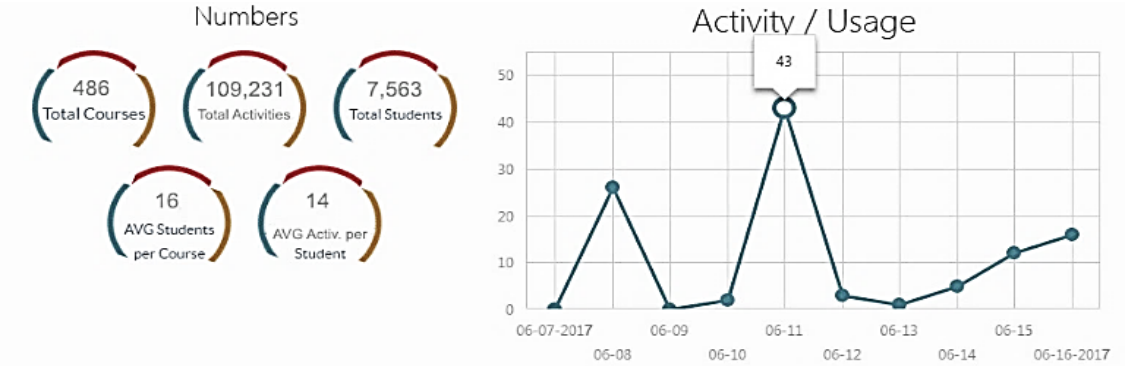


Figure 4. Statistics in Lentera

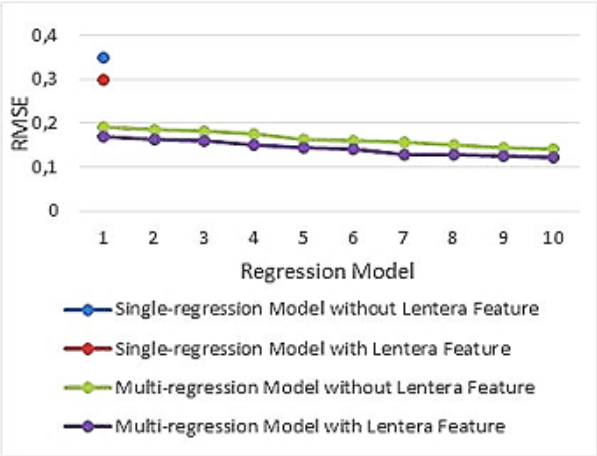


Figure 5. The graphic of regression model vs RMSE

4. CONCLUSION

In this research, multi-regression model to forecast the performance of university student was implemented. According to the testing result, multi-regression model performs better in explaining dependent variables 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] Santoso. L. W., "Analysis of the impact of information technology investments – a survey of indonesian universities," *ARPN JEAS*, vol. 9, no. 12, pp. 2404-2410, 2014.
[2] Baker R. and Inventado. P, "Educational data mining and learning analytics," *Learning Analytics,* pp. 61-75, 2014.

[3] Anderson J. R., Boyle C. F., and Reiser B. J., "Intelligent tutoring systems," *Science,* vol. 228, no. 4698, pp. 456-462, 1985.
[4] Mjhool A. Y., Alhilali A. and H, Al-Augby S, "A Proposed architecture of big educational data using hadoop at the university of kufa," *International Journal of Electrical and Computer Engineering,* vol. 9, no. 6, pp. 4970-4978, 2019.
[5] Romero C., and Ventura S., "Educational data mining: A review of the state of the art," *Trans. Sys. Man Cyber Part C*, vol. 40, no. 6, pp. 601-618, 2010.
[6] Santoso L. W., Yulia, "The analysis of student performance using data mining," *Advances in Intelligent Systems and Computational Sciences,* pp. 559-573, 2019.
[7] Santoso L. W, "Early warning system for academic using data mining," *Fourth International Conference on Advances in Computing, Communication & Automation,* pp. 1-4, 2019.
[8] Santoso L. W., "Data warehouse with big data technology for higher education," *Procedia Computer Science,* vol. 124, no. 1, pp. 93-99, 2017.
[9] Barber R., Sharkey M., "Course correction: Using analytics to predict course success," *International Conference on Learning Analytics and Knowledge,* pp. 259-262, 2012.
[10] Wang J., and Karypis G., "Harmony: Efficiently mining the best rules for classification," *Data Mining Conference,* pp. 205-216, 2005.
[11] Han J., Pei J., Yin Y., "Mining frequent patterns without candidate generation," *ACM SIGMOD Int'l Conf. on Management of Data,* vol. 29, no. 2, pp. 1-12, 2000.
[12] Fradkin D. and Morchen F., "Mining sequential patterns for classification," *Knowl. Inf. Syst,* vol. 45, no. 3, pp. 731-749, 2015.
[13] Agrawal R., Golshan B., and Papalexakis E. E., "Toward data-driven design of educational courses: A feasibility Study," *Journal of Educational Data Mining (JEDM)*, vol. 8, no. 1, pp. 1-21, 2016.
[14] Jittawiriyanukoon C., "Proposed classification for elearning data analytics with MOA," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 3569-3575, 2019.
[15] Andayani S., et al, "Decision-making model for student assessment by unifying numerical and linguistic data," *International Journal of Electrical and Computer Engineering,* vol. 7, no. 1, pp. 363-373, 2017.
[16] Santoso L. W., "ITIL service management model for e-learning," *Journal of Adv Research in Dynamical & Control Systems,* vol. 11, no. 6, pp. 190-197, 2019.
[17] Quadri M. N, and Kalyankar N. V., "Drop out feature of student data for academic performance using decision tree techniques," *Glob. J. Comput. Sci. Technology*, vol. 10, no. 2, pp. 2-5, 2010.
[18] Dekker G. W., Pechenizkiy M., and Vleeshouwers J. M., "Prediction student drop out: A case study," *2nd International Conference on Educational Data Mining,* pp. 41-50, 2009.
[19] Barber R., and Sharkey M., "Course correction: Using analytics to predict course success," *2nd International Conference on Learning Analytics and Knowledge,* pp. 259-262, 2012.
[20] Leitner P., Khalil M., and Ebner M., "Learning analytics in higher education ð a literature review," *Learning Analytics: Fundaments, Applications, and Trends. Springer International Publishing,* pp. 1-23, 2017.
[21] Lee Y., and Cho J., "An intelligent course recommendation system," *Smart CR,* vol. 1, no. 1, pp. 69-84, 2011.
[22] Yunianta A., et al., "Solving the complexity of heterogeneity data on learning environment using ontology, *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 13, no. 1, pp. 341-348, 2015,
[23] Ventura C. S., Espejo P. G., and Hervás C., "Data mining algorithms to classify students," *1st Int. Conf. on Educational Data Mining. Montreal,* pp. 187-191, 2008.
[24] Thai-Nghe Ng, et al., "Factorization techniques for predicting student performance. educational recommender systems and technologies: Practices and challenges," *OIGI Global,* pp. 129-153, 2011.
[25] 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 educational data mining,* vol. 32, 2006.
[26] Lamani A., et al., "Data mining techniques application for prediction in OLAP Cube," *International Journal of Electrical and Computer Engineering,* vol. 9, no. 3, pp. 2094-2102, 2019.
[27] Daniel B., "Big data and analytics in higher education: Opportunities and challenges," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 904-920, 2014.
[28] Elbadrawy A., Studham R. S., and Karypis G., "Personalized multi-regression models for predicting students' performance in course activities," *International Conference on Learning Analytics and Knowledge,* pp. 103-107, 2015.
[29] Agarwal D., and Chen B. C., "Regression-based latent factor models," *ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* pp. 19-27, 2009.