Student Next Assignment Submission Prediction Using a Machine Learning Approach

Y. K. Salal[1(✉)], M. Hussain[2], and T. Paraskevi[3]

[1] South Ural State University, 76 Lenina Avenue, Chelyabinsk 454080, Russian Federation
yasskhudheirsalal@gmail.com
[2] Shanghai University, 266 Jufengyuan Street, Shanghai 200444, China
[3] University of Piraeus, 18 M. Karaoli & A. Dimitriou Street, 18534 Piraeus, Greece

Abstract. The web-based learning platform provides quality education nowadays, but assignment submission is a critical issue in the e-learning system. Therefore, to investigate assignment submission of the student in advance before the end of course is an important problem. The assignment submission prediction is the advantage of the e-learning system because it allows the instructor to find students' problems on time. Additionally, online learning mostly depends on demographic characteristics such as region, age, and education level. This study uses machine learning (ML) methods to detect students who do not submit assignments on time and then also find which demographic factors affects online assignment submission. The data is publicly available and was collected from an open university of U.K. The result shows that Random Forest is an optimal option for predicting students who do not submit an assignment on time. We have also found that Gender, Student Credit, Final result, Total clicks, Score are strong predictors for student assignment.

Keywords: Machine learning · E-learning · Student assignment · Feature extraction · Classification

1 Introduction

Online learning allows all ages, educational levels to participate in learning activities. It is also helpful in many emergency situations such as Coronavirus disease (COVID-19). In online learning, the student assignment submission rate affects student performance in traditional education as well as the e-learning system.

Several studies have been conducted on other problems of e-learning systems such as student dropout, student performance; but there is little research work on student assignment submission prediction and impact of student demographic characteristics on assignment submission. Cai et al. [1] shows how demographic characteristics of online learning students affect the learning outcome. Hooshyar et al. [2] devised a novel algorithm to predict the students' performance with learning difficulties by utilizing assignment submission behavior via procrastination behavior (termed as PPP). PPP examined the behavioral patterns of the students besides considering the late or non-submission

A. A. Radionov and V. R. Gasiyarov (Eds.): RusAutoConf 2020, LNEE 729, pp. 383–393, 2021.
https://doi.org/10.1007/978-3-030-71119-1_38

of assignments. Herodotou et al. [3] investigated whether predictive learning analytics (PLA) might be utilized to help at-risk students. Fernández-Alonso et al. [4] analyzed homework assignment strategies in school that affected the performance of the students. They measured academic performance of 26,543 Spanish adolescents in four subjects-Citizenship, Science, Mathematics, and Spanish. Wakelam et al. [5] employed machine learning and data mining strategies to assess tutorial/lecture attendance, interim assignments to detect potential at-risk students of failure or withdrawal. Songsirisak et al. [6] investigated the impact of homework assignments on students' learning. Abdillah et al. [7] suggested some of the students' participation strategies based on blog and e-learning. These activities help students to learn independently. In the e-learning environment, students can discuss with their colleagues, give comments, submit their assignments and participate in learning scenarios. Raitman et al. [8] addressed the issues related to social security in collaborative learning environments. They conducted a case study in wiki platform in the tertiary education environment as a way of online collaboration. In the e-learning environment, where the assignment submissions are in electronic form, basic security requirements such as the availability, confidentiality and integrity must be assured. You et al. [9] examined the effect of academic procrastination in a learning management system (LMS). Procrastination involves late submission of assignments and delays in weekly scheduled learning which can be detected from the log data. Santoso et al. [10] evaluated a learning management system by utilizing User Experience Questionnaire (UEQ). They used a LMS termed as Student Centered e-Learning Environment (SCELE). The findings demonstrated that submission of assignments using SCELE was an easier task (15%) and 21.4% learners agreed that the assignment submission tool was a key feature in the LMS.

In our study, we predict the next assignment submission through various machine learning techniques. Assignment submission is vital in an e-learning environment. Our framework is based on demographic factors such as region, age, and education level for such prediction. We have applied the K-Nearest Neighbor (KNN) classifier for both training and testing of data. KNN algorithm exhibits an accuracy of 0.78. We have applied logistic regression for training the dataset and also utilized Random Forest (RF) classifier on the test data. RF classification method yields an accuracy of 0.93 by outperforming the KNN method. This result will help the instructors to find the reason for a student; who does not submit an assignment on time. By examining the student assignment submission, it also allows the teachers to discover students who have difficulty in the course and give feedback on time. Additionally, it will help instructors to give guidance to students in their assignment.

2 Problem Description

In this study, we use student activities data from an online learning system (Virtual Learning Environment (VLE)). The VLE is an important environment to improve the student's learning. It allows the instructor to design and deliver their course online. In the e-learning system, the students are enrolled and read material anytime from any place. The problem of detecting the student; who does not submit an assignment on time is an important issue in VLE. We used data visualization and machine learning techniques to

investigate this problem. Finally, we find the factor which affects the student assignment submission. The mathematical explanation of our problem is as follows.

$$T = \{x_i, y_i\}_{i=1}^{N}, \tag{1}$$

where, x_i is an N-dimensional input vector with input features, y_i is a vector of the target class; T is the training set in the study. The features include final results of the students, total number of clicks on the VLE activities, age, num previous attempts, highest education, region, gender, disability, student credit, score, date submitted. N represents the number of students taken into consideration ($N = 388$). $y_i \in [1, 0]$, y_i is set to 1 if the student submits the assignment, otherwise it is 0.

3 Materials and Methods

In this study, we investigate the student assignment submission problem through various machine learning techniques. The main steps of this study are as follows.

3.1 Data Description and Study Context

In the current study, we have analyzed the social science course data which was attended by Open University students. The Open University UK delivered this course through VLE. Using VLE students enrolled into the course and performed different activities during the course. This study is based on 388 students' data after they completed the course assignment using VLE. This data is available online for research purposes. The period of study of those students was from 2013 to 2014.

In this study, data shows the students' performance, behaviour during completing assignments. This information is used to investigate the reason; why students does not submit course assignments and then find the factors which affects the student assignment submission process.

3.2 Current Study Methods

The current study framework is shown in Fig. 1.

3.3 Feature Extraction

In feature extraction, we extract the relevant features to predict the next assignment submission. We have used Ms excel to extract the relevant feature to this problem. The details of these features are shown in Table 1.

3.4 Data Discretization and Label Assignment

In data discretization, we convert the numerical value of the dataset into categorical. For modelling our problem in machine learning, we consider the binary variables that is 0 to predict students who do not submit an assignment on time and 1 to predict students who submit an assignment on time.

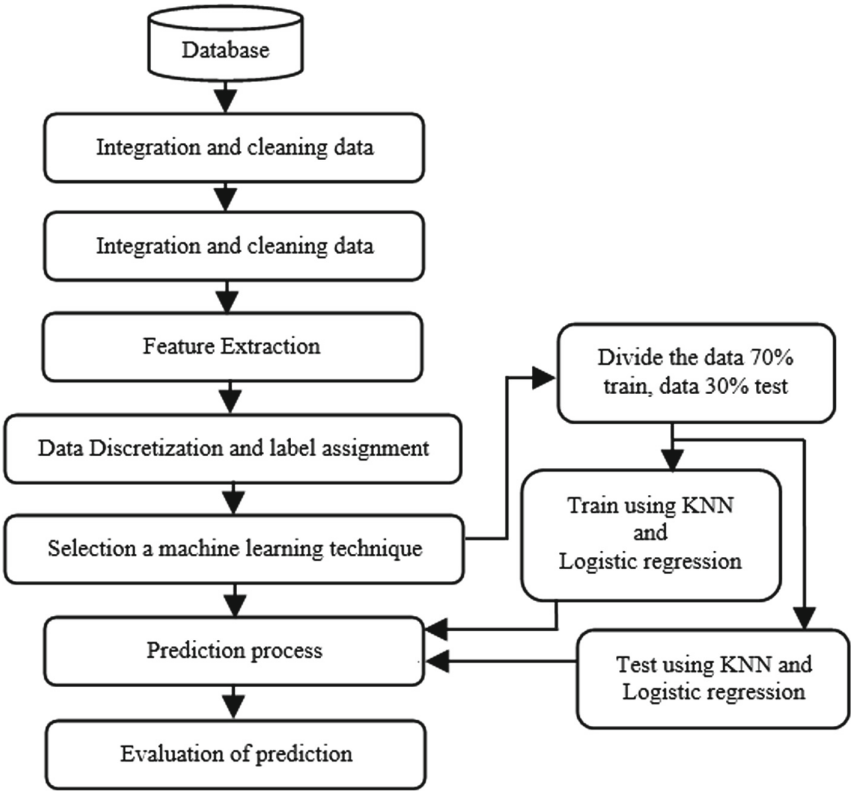

Fig. 1. Data flow of the current study.

Table 1. Features in the dataset with their description.

Features	Description	Features	Description
Student ID	The student's identification code	Region	The region from where the student resides
Final_result	The final results of the student whether the student is passed/failed/withdrawn	disability	Whether the student is disabled
Num_of_prev_attempts	The total number of earlier attempts	studied_credits	The number of credits taken by the student
Age	The age of the student	score	The score attained by the student
highest_education	The highest degree of the student	sum_click	The number of clicks in VLE
Gender	The gender of the student	date_submitted	The date of submission the assignment

3.5 Classification Technique

Random Forest (RF) [11]: Random Forest consists of a large number of individual decision trees that operate as an ensemble. Each individual decision tree in the random forest predicts and class with the most votes is the prediction of the model. The prime concept of random forest is simple but powerful- it is the wisdom of the crowds. Uncorrelated models can exhibit ensemble predictions that are higher in accuracy than the individual predictions. Hence, a large number of uncorrelated trees operating as a committee can outperform any of the individual constituent models.

K-Nearest Neighbor (KNN) [12, 13]: KNN is an easy to implement simple algorithm that can be used to solve both regression and classification problems. The classifier assumes that related things exist in close proximity. Similar things exist near to each other. One of the advantages of KNN is that there is no need to tune additional parameters, make assumptions or build a model. The algorithm is versatile as it can be used for search, classification as well as regression. The disadvantage of the algorithm is that as the predictors increase, the slower the performance of the algorithm. To select the right K is the crucial thing. The algorithm runs several times to choose K such a way that it reduces the number of errors.

Logistic Regression [14]: Logistic Regression is used in many biological and social science applications. It is used when the target variable is categorical. There are different types of logistic regression. They are binary, multinomial and ordinal logistic regression. In the binary version, the categorical variable has only 2 responses. In multinomial logistic regression, there are three or more categories without any particular order. In the ordinal version, there are 3 or more categories with ordering. A threshold can be set to predict which class a data belongs to. Decision boundaries can be non-linear or linear. Polynomial order can be enhanced to achieve complex decision boundary.

4 Experiments and Results

In this paper, we have classified the students in two groups, students who submit assignments and other students; who do not submit assignments. If the student data can be used to predict group memberships, then the e-learning university can customize their feedback for individual prospective students. This is a classification problem that is given a dataset with a pre-defined label. We need to build a model to predict the class of new students. The target field called assignment submission has two possible values corresponding to two students' groups (assignment submitted, assignment not submitted).

Our objective is to build the classifier using the student data to predict student class and find which features affect more the student class. We use different types of classification algorithms. This section contains the following sub portions.

4.1 Data Description and Visualization

With help of data visualization, we can try to understand the pattern, trend, and correlation between the variables.

Figure 2a shows how a student; who submit and not submit assignment related with student qualification. Figure 2a shows that the assignment submission ratio in A level qualified student is high as compared to other students. Mostly the post-graduation student does not submit more assignments as compared to other students.

Figure 2b shows the relation between student assignments submission state with students' age. The result shows that students among the age group 0–35 cannot submit more assignments as compared to other students. One possible reason for that, students in this age are busy in other activities such as collage, university and sport etc.

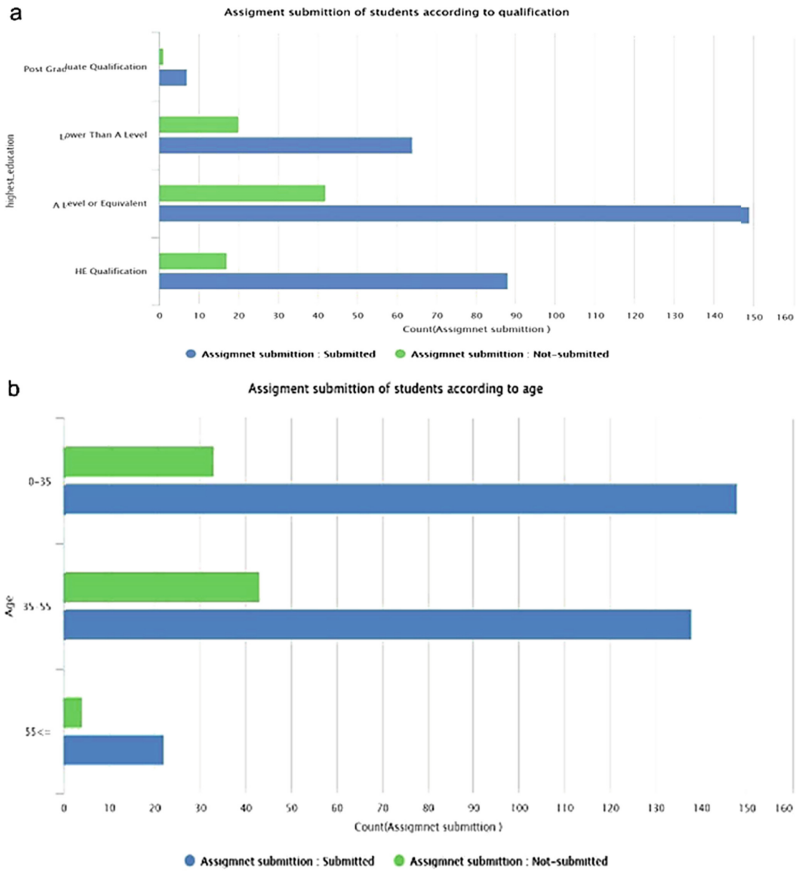

Fig. 2. Visualization the students assignment state with (a) students qualification; (b) students age.

Figure 3 shows the relationships between the target variable and region. The results show that the assignment submission rate in the East Anglian region is high as compared to other regions. The student assignment not submission ratio in North and Ireland region is high as compared to another region.

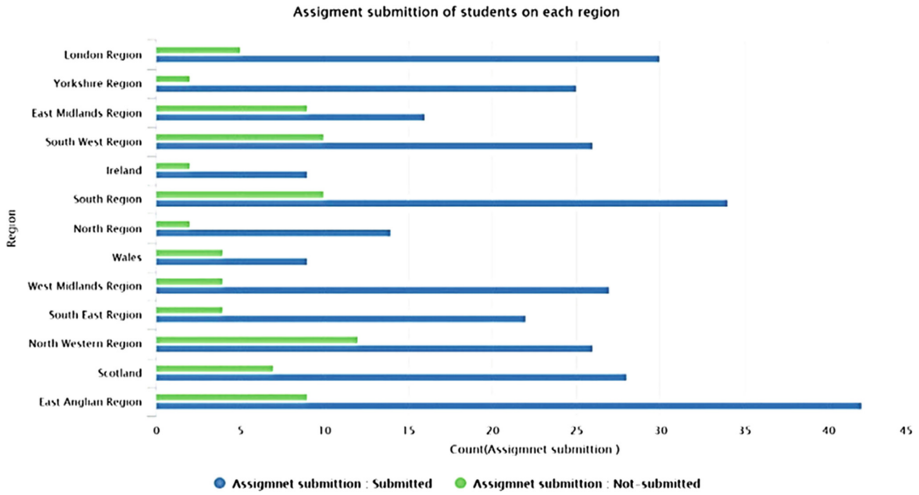

Fig. 3. Visualization the student's assignment state with student's region.

4.2 Design of the Predictive Models and Experimental Setups

This study uses an analytical data set of an open university. In this study, 388 students enrolled in the course. The students completed different assignments during the course. The sklearn module of python is used to perform different tasks such as visualized data; description of data and build machine learning models. We build different machine learning models because these models are largely used in education to find the pattern of the students. We tune different parameters of machine learning algorithms to get better results. The input variables of the current study are as shown in the Table 1 and output variables are submitted or not submitted. We randomly divide the student's data into 70% training and 30% testing portions so that it reduces the problem of overfitting [15]. We use the first assignment data of students. Some features in our dataset are categorical; therefore we use the "LabelEncoder" module of sklearn to convert these features in numeric because some ML algorithms can handle the categorical attributes. The random forest achieves the best performance on the following parameter such as max_depth = 3 and n_estimators = 10.

The logistic regression algorithms setting are random state = 0. The KNN model achieves good accuracy on setting k = 1.

4.3 Performance Measurements

To select the best ML algorithms for current study, we use accuracy performance parameters and confusion matrix. Accuracy shows correctly classified items in the test data set. High accuracy indicates that model performance is good in predicting students; who do not submit assignments.

$$\text{Accuracy} = \text{TF} + \text{TN}/(\text{TF} + \text{FN} + \text{FP} + \text{TN})$$

Note: True positive (TP); True negative (TN); False Negative (FN); False Positive (FP).

4.4 Results

In this section, we want to find which ML algorithms are suitable to detect students; who do not submit and submit assignments based on student data during the course. Next, we want to find which age group and specific region does not submit an assignment on time. This will help the instructor to customize their feedback and send mail to those students. We conduct experiments to achieve these goals.

To find appropriate classifiers to predict students; who do not submit an assignment during the course; we build and compare the performance of the following classifiers.

We use training data to train the logistic regression classifier. Then, we test RF classifiers on test data. The logistic regression received accuracy = 0.88. The results are shown in Table 2. There are a total 28 students who belong to class "Not-submitted". Out of 28 students, 15 students of class Not-submitted were successfully classified and 13 students were wrongly classified by classifier.

Table 2. Confusion matrix of the Logistic Regression classifier.

Predicted/Actual	Not-submitted	Submitted	Class precision
Not-submitted	TP = 15	FN = 13	1.0
Submitted	FP = 0	TN = 89	0.87
Class Recall	0.70	0.93	

Next, we train and test KNN classifier. It obtains 0.78 accuracy. The confusion matrix of KNN classifier is shown in Table 3. The result shows that out of 28 not-submitted students, 12 students are successfully classified and 16 students are wrongly classified. One challenging task in KNN classifier is to find the appropriate K value because finding the optimal value of K impacts the performance of the model. This study tries different values of K that is up to 10. We use a scleral module to find appropriate K value which is shown in Fig. 4.

Finally, we estimate the performance of the RF classifier to predict students who do not complete assignments on time. RF classifier yields high accuracy = 0.93 using student data and correctly predicts 20 students out of 28 students who do not submit assignments.

The result shows that RF classifier predicts the student who does not submit assignment with high accuracy as compared to the other model of the study (Table 4). The RF classifier decision tree is shown in the Fig. 5.

4.5 Discussion

There are two main goals of this study, First find the appropriate ML model which predicts not submitted assignment students with high accuracy and second find which input features greatly affect the classifier.

Table 3. Confusion matrix of the KNN classifier.

Predicted/Actual	Not-submitted	Submitted	Class precision
Not-submitted	TP = 12	FN = 16	0.57
Submitted	FP = 9	TN = 80	0.83
Class Recall	0.43	0.90	

Table 4. Confusion matrix of the RF classifier.

Predicted/Actual	Not-submitted	Submitted	Class precision
Not-submitted	TP = 20	FN = 8	1
Submitted	FP = 0	TN = 89	0.92
Class Recall	0.71	1	

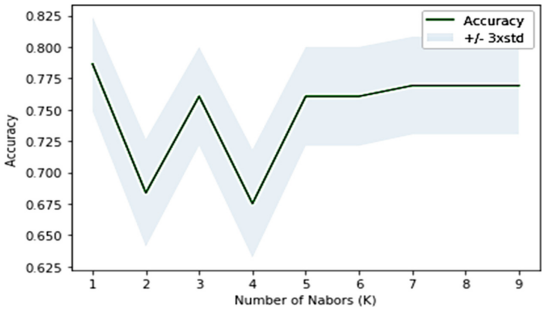

Fig. 4. Find appropriate K value for KNN classifiers.

Research Question 1: What are the appropriate ML classifiers for predicting the students who are not going to submit an assignment on time?

In order to find which ML technique is appropriate for this problem; we explore first question. In machine learning, the performance of learning models depends on classification data. The results show that the performance of the decision tree model (RF classifier) predicts not-submitted assignment class with high accuracy because DT classifiers are appropriate for categorical data as compared to other models of this study. Furthermore, DT classifiers split the difficult task into simple classification tasks and use a bagging feature to predict the not-submitted student with high accuracy.

- The performance of KNN and logistic regression is not good as compared to RF classifiers. One possible reason is that KNN needs appropriate K value which is a challenging task for the researchers and also affects the performance of models
- Research Question 2: What are the demographic factors that affect the student assignment submission?

In order to find how the student input features are related to student assignment state, we explored the second question. We plot trees using RF classifiers which are shown in Fig. 5. We got some important results during plotting the decision tree using RF which is interpreted as follows.

The Fig. 5 shows that students Gender, Student ID, Student Credit, Final result, Total clicks, Score play an important role in predicting the students who do not submit assignments during course because these features are shown many times in the decision tree. Other input variables are not shown in the plot which indicates that these variables are not vital for model prediction.

The graph shows that total click is an important predictor of this study because number clicks show engagement of students in solving the exercise, greater the number clicks they have a greater chance to submit the assignments. Furthermore, the graph shows that high score students have a greater chance to submit assignments as compared to low score students.

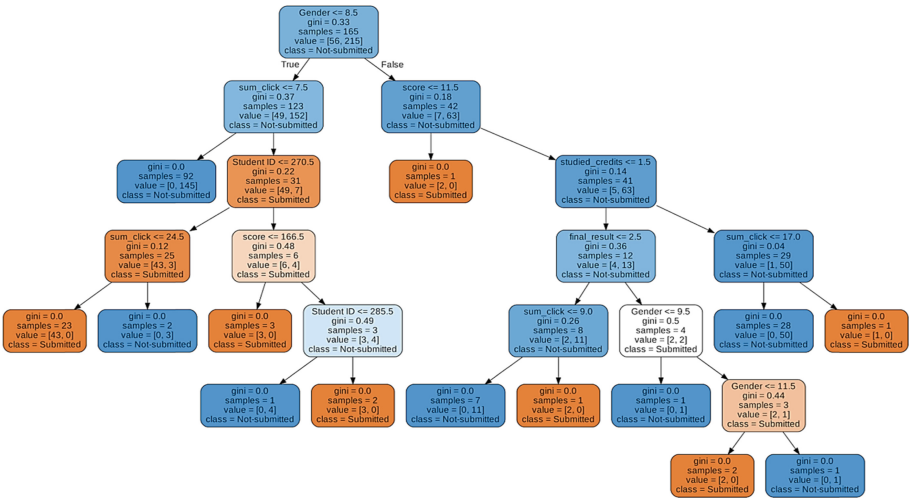

Fig. 5. Plot the decision tree using RF classifier.

5 Conclusions and Future Works

newpageThe study successfully predicts the next assignment submission of learners in an e-learning environment. The assignment submission is crucial in the learning management system. Different machine learning classifiers are employed for such predictions. The Random Forest classifier proves its efficacy with an accuracy of 0.93. The results also show that students Gender, Student ID, Student Credit, Final result, Total clicks, Score are important factors which affect the student assignment submission.

Data Availability. The current study data are publicly available online (https://analyse.kmi.open.ac.uk/open_dataset) for research purposes. Ethical clearance was granted by the Open University, UK. No participants' personal information (e.g., name or address) was included in this study.

References

1. Cai, Z., et al.: Gender and attitudes toward technology use: a meta-analysis. Comput. Educ. **105**, 1–3 (2017). https://doi.org/10.1016/j.compedu.2016.11.003
2. Hooshyar, D., Pedaste, M., Yang, Y.: Mining educational data to predict students' performance through procrastination behavior. Entropy **22**(1), 12 (2020). https://doi.org/10.3390/e22010012
3. Herodotou, C., Rienties, B., Boroowa, A., et al.: A large-scale implementation of predictive learning analytics in higher education: the teachers role and perspective. Educ. Tech. Res. Dev. **67**(5), 1273–1306 (2019)
4. Fernandez-Alonso, R., Alvarez-Diaz, M., Suarez-Alvarez, J., et al.: Students' achievement and homework assignment strategies. Front. Psychol. **8**, 286 (2017). https://doi.org/10.3389/fpsyg.2017.00286
5. Wakelam, E., Jefferies, A., Davey, N., et al.: The potential for student performance prediction in small cohorts with minimal available attributes. Br. J. Edu. Technol. **51**(2), 347–370 (2020). https://doi.org/10.1111/bjet.12836
6. Songsirisak, P., Jitpranee, J.: Impact of homework assignment on students' learning. J. Educ. Naresuan Univ. **21**(2), 1–9 (2019)
7. Abdillah, L.A.: Students learning center strategy based on e-learning and blogs. arXiv preprint arXiv:1307.7202 (2013)
8. Raitman, R., Ngo, L., et al.: Security in the online e-learning environment. In: Fifth IEEE International Conference on Advanced Learning Technologies, pp. 702–706 (2005)
9. You, J.W.: Examining the effect of academic procrastination on achievement using LMS data in e-learning. J. Educ. Technol. Soc. **18**(3), 64–74 (2015)
10. Santoso, H.B., et al.: Measuring user experience of the student-centered e-learning environment. J. Educators Online **13**(1), 58–79 (2016)
11. Liaw, A., Wiener, M.: Classification and regression by random Forest. R News **2**(3), 18–22 (2002)
12. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
13. Mukesh, K., Salal, Y.K.: Systematic review of predicting student's performance in academics. Int. J. Eng. Adv. Technol. **8**(3), 54–61 (2019)
14. Kleinbaum, D.G., et al.: Logistic Regression. Springer-Verlag, New York (2002)
15. Abdullaev, S.M., Salal, Y.K.: Economic deterministic ensemble classifiers with probabilistic output using for robust quantification: study of unbalanced educational datasets. In: International Scientific and Practical Conference on Digital Economy, vol. 105, pp. 658–665 (2019)