JOIV : Int. J. Inform. Visualization, 5(4) - December 2021 347-353

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Early Dropout Prediction in Online Learning of University using Machine Learning

Hee Sun Park [a], Seong Joon Yoo [a,*]

[a] *Department of Computer Science, Sejong University, 209 Neungdong-ro Gwanging-gu , Korea ,Seoul, 05006, South Korea*
Corresponding author: [*]sjyoo@sejong.ac.kr

***Abstract*— Recently, most universities plan to open or open online learning courses, but the problem of dropout of online learning is still a problem for universities. Online learning has the advantage of being able to receive education anytime, anywhere, but it is true that the dropout rate is higher than offline classes because you have to manage and control your own study time without the help of a professor or manager. Therefore, it is very important for professors and managers to support students in a timely act to avoid the risk of dropout of university online classes. This study used the access log data recorded in the Learning Management System (LMS) and the learner's statistical information and calculated data, and aims to present predictive algorithms suitable for online learning dropout early prediction systems at universities. This study features a 7-year online learning history log data recorded in the Cyber University LMS system to overcome the data count limitations of existing studies and predict the risk of drop-out during the learning period. The characteristics of the data you utilized were used to validate the availability of predictive models by applying learner statistical information, number of system connections, number of lectures, previous semester grade data, machine learning based decision tree, arbitrary forest (RF), support vector machine (SVM) and deep learning (DNN). Studies show that random forest (RF) algorithms have the best prediction and performance, and deep learning algorithms also apply to learning management (LMS) systems.**

***Keywords*— Dropout prediction; online learning; machine learning; deep learning.**

Manuscript received 8 Feb. 2021; revised 22 Mar. 2021; accepted 5 Apr. 2021. Date of publication 31 Dec. 2021.

I. INTRODUCTION

Online education is a good way to help learners learn new knowledge and get a degree without any time or space constraints. Although online edcation is receiving much attention in universities because most universities have recently opened or plan to open online edcation courses, there is a problem that the dropout rate is higher than offline edcation because learners' time management and supervised edcation activities are not inherent in online education. In the case of learners who participate in online edcation with goals such as self-development and degree acquisition, dropping out will lose time and economic power, which eventually leads to suspension of self-development and failure to obtain a degree. Therefore, lowering the dropout rate requires active intervention by teachers or managers if learners are likely to drop out, and developing early dropout prediction services is an important challenge for universities. Therefore, it is true that not much research has been conducted on the feasibility of predictive models to predict such an early dropout..

This study aims to develop an effective prediction model for applying to effective dropout prediction systems that predict the likelihood of dropping out of these online learning in advance and inform teachers or managers.

There have been numerous studies related to the dropout of online learning. However, this study has the following differences from previous studies.

First, it is meaningful in studying models suitable for application to online learning at universities. Recently, dropout prediction analysis using online learning history data has been actively conducted, but most of them are based on MOOCs online courses, making it difficult to apply to online learning at universities.

Second, this study is characterized by studying a model that can predict early dropouts on a weekly basis. Most predictive studies of online learning dropout have problems that cannot predict dropout for learners in progress as they utilize learner surveys after learning is completed and data stored in the learning system recently. In this work, we can inform universities and professors to early dropout features early in the learning period, as weekly data analysis rather than data

after completion of learning allows early dropout of dropout features.

Third, this study overcame the difficulty of generalization due to small amounts of data. It utilizes actual learning history data collected in the learning management system(LMS) of online universities in Korea for more than a decade.

Previous studies related to this study can be divided into factor analysis domains that have a significant impact on online learning and predictive studies to prevent early dropouts as follows

A. A Study on the Dropout features of Online Learning

According to previous studies, that since online learning takes place in spatially and spatially separated situations, it has emphasized self-directed learning capabilities, among other things, the importance of learning analysis data, which is calculated from learners' learning and learning records [1][2]. There are studies explaining that learning activity data (free bulletin boards, lecture boards, learning material boards, discussion boards, and content module access rates) have a significant impact on learning performance, and that detailed factors (major, gender, age, grade, academic, occupation) also have a greater impact on learning performance[3]. There are studies that show that predictive power increases the most learning and predictive power when using data from learner individuals, learning environments, and learning processes comprehensively [4].

B. A Study on the Prediction of Dropout of Online learning

Recently, research on predictive models has been actively conducted around MOOC lectures to inform the need for intervention for students at risk of dropping out. There is a study that develops a weekly temporal dropout prediction model and proposes a model that provides information about personalized interventions using individual dropout probabilities [5]. [6] has applied machine learning methods to predict the learning performance of large-scale open online courses (MOOCs) and shows that there is a strong correlation between learners' click behaviors and learning outcomes. We also explain that among various machine learning methods, random forests show the best performance for prediction problems. A session-by-session dropout prediction study of MOOC courses defines learning and predictive units as sessions and utilizes Long Short-Term Memory (LSTM) and d Gated Current Unit (GRU) algorithms shows that LSTM models perform up to 12.2% better than GRU models (Based on AUC) [7].

II. MATERIAL AND METHOD

For the development of predictive models, it is important to select the data to analyze and the features of the data, and the algorithms to be used in the predictive model must also be chosen for the purpose. Once the prediction model is created, it is necessary to go through the steps of evaluating how accurate the prediction will be in actual situations. These flows are as follows:

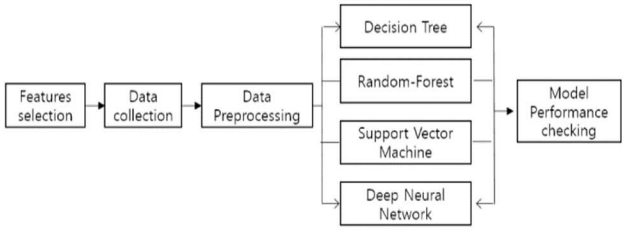

Fig. 1 Flow Chart of Study

1) Data Collection: Using the log data in Learning management System Of Cyber university (2012.03.01~ 2019.12.01)

2) Features Selection: The week information, User information, Previous learning information, The access log , The Activities in classes , The Status Dropout

3) Data Preprocessing: Using Oversampling, Normalization, one-hot encoding technique, etc

4) Algorithm Selection : Decision Tree, Random-Forest(RF), Support Vector machine(SVM), Deep Neural Network(DNN)

5) Evaluation: Accuracy, Recall(Sensitivity), Precision, F-measure, ROC curve

A. Data Collection

This study utilized 98,685 student statistical information from March 2012 to December 2019 and 1,480,275 log data stored in the learning management system for online learning. The student's statistical information, such as the number of enrollments, scholarship status, age, and course registration, was collected in the academic administration management system(ADS), and the weekly access records and learning activity records were collected in the learning management system(LMS)

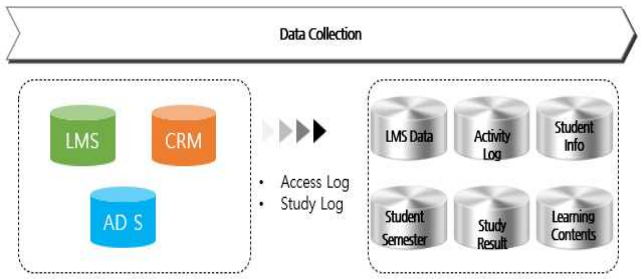

Fig. 2 Selection of Data

B. Features Selection

The conditions for online learning early dropouts vary according to the objectives of the prediction model, depending on whether the course was completed, whether re-enrollment occurred in the following semester, grades, etc. This study emphasizes early dropouts in the online courses of universities, and so whether a learner re-enrolled in the following semester was set as a condition of dropping out. In order to predict dropouts as a course unit of courses such as MOOC, it was deemed reasonable to set such conditions as whether the course was completed and grades as conditions for dropping out.

It was deemed important to select usage features which were used as both conditions for dropouts and prediction analysis. In this study, previous studies on the analysis of online learning early dropouts were reviewed and features checked which were deemed to have a high degree of influence on dropouts, and the finalized selection was of features able to be used and gathered from a universitiy online learning LMS system. The features utilized in this study are displayed in Table 1.

TABLE I
FEATURES AND DESCRIPTION

No	Features	Description
1	Analytical Units	Week (1 week to 15 week)
2	Student information	Semesters of enrolled
3		Scholarship
4		Age group
5		Gender(male or female)
6		Previous Degree
7		Address(Domestic/Overseas)
8		Transferring schools
9		Multiple majors selected
10	Online learning information	Number of classes currently enrolled information Total access
11	Activity information	Total access up to the current week (lms)
12		Number of days of access up to the current week (lms)
13		Total Learning time
14		Avg Learning time(Total learning time / number of classes)
15		Average number of class notification notices read (per class)
16		Average number of task submissions(per class)
17		Average number of bulletin board activities(per class)
18		Average number of examinations(per class)
19	Previous Online learning Information	Total number of online courses up to the previous semester
20		Number of online courses for the previous semester
21		Overall grade of previous online training courses
22		grade of previous online training courses
23	Next semester Infomation	1: Dropout 0: Non-Dropout

In this work, we use a total of 23 data features using a total of 22 input features and re-enrollment items for early dropout judgment. The main features selected are described in the following:

- Week : It is unit information of analysis and prediction. In the case of universities, information from 1 to 15 weeks was used because learning takes place over a semester from 1 to 15 weeks
- Scholarship : Whether the student has received a scholarship at the time of enrollment for this semester
- Previous Degree : Degree information before entering college (college or high school)
- Transferring schools : Whether it is a freshman or not, transfer schools may have a strong willingness to study or have a strong resentment to online learning.
- Multiple majors selected : Whether you have completed a double major or a minor other than a major, and if you choose a double major, you have a strong willingness to study.
- Total access up to the current week (lms): The number of online learning systems up to the current week on which they are based. The more times, the less likely it

- *Overall grade of previous online training courses : Higher grades up to the previous semester are less likely to deviate; lower grades are more likely to deviate*

- *Status Dropout: As universities often calculate dropping out based on the withdrawal of enrollment next semester, whether to re-enroll next semester has been selected as a dependent variable of the model.*

Many previous studies used whether or not to deviate from the subject's grades as a criterion for determining whether to re-enroll the next semester, which is easy to apply to universities, as a criterion for determining whether to re-enrollment.

By examining the correlation between the selected features, we confirmed the correlation between the 22 features selected to exclude features that are not related to dropout from analysis and the re-registration value that determines whether to drop out.

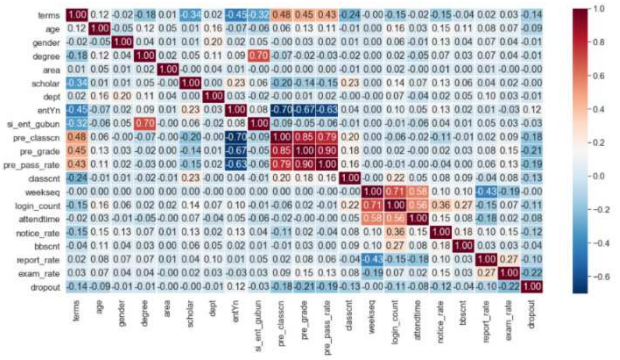

Fig. 3 Confirming the correlation between features

In this study, there were no features excluded from the correlation analysis because 22 features with high correlation were selected through previous studies on feature extraction.

C. Data Pre-processing

The data pre-processing process of making collected data suitable for analysis with machine learning algorithms is essential. In the study, the following methods were used in the data preprocessing process:

1) *Converting Data to Numeric :* Convert string data such as scholarships, gender, previous degrees, addresses (domestic/foreign), multiple majors, and enrollment status data for the next semester into numbers.

2) Removal of incomplete data (missing) : Data of students who did not have information for the previous semester, such as returning to school or re-entry, will be deleted.

3) Removal Noise Data : Previous degree information, address information, sometimes missing values, so data is deleted

4) Removal contradictory data: such as when a male social security number starts with 2, etc.

5) *Resolving data imbalance :* Oversampling of dropout label data

D. Algorithm Selection

For early dropout predictions, it is most important to choose the best prediction model(algorithm). This work utilizes machine learning, an artificial intelligence algorithm that is effective for processing large amounts of data, referring to prior research on big data to implement optimal prediction models. In particular, deep learning algorithms conducted experiments to increase performance and accuracy through hyperparameter optimization. Each algorithm has the following characteristics:

1) *Decision Tree:* The Algorithm is widely used for classification and regression problems, especially which has the advantage of clearly understanding how the algorithms predicted with visualization, while having the disadvantage of performance degradation due to oversampling of training data.

Random-Forest (RF): The algorithm is the most commonly used model for classification and regression in decision tree models that solve the performance degradation problem caused by oversampling of training data. However, Random Forest uses more memory than linear models and has a disadvantage of slow training and prediction.

Support Vector Machine (SVM): The algorithm has the advantage of operating with a small number of data characteristics, but it has the disadvantage of having speed and memory problems with increasing sample size, and it is difficult to understand how predictions were determined in difficult analyses.

Deep Neural Network (DNN): The algorithm is Designed for deep hidden layers of artificial neural networks, it is effective in distinguishing key content or features in complex materials in a way that most closely resembles human mindsets. In particular, good performance is shown in image analysis and natural language analysis, and good performance can be expected in high data volumes. However, hidden layer composition and hyperparameter tuning are very important because of the disadvantage of overfitting problems and time-consuming for learning.

E. Evaluation

Predictive models generated from training data should verify the accuracy of the predictions compared to the verification data, and predictions from the untrained data should be verified by comprehensively reflecting not only accuracy but also recall, precision, etc. Therefore, in this work, we demonstrate through the F-measurements of the ROC curve and the AUC values. The evaluation methods used in this study are as follows Table II

TABLE II
EVALUATION AND DESCRIPTION

No	Evaluation	Description
1	Accuracy	The proportion of correct predicted data (TP + TN)/(TP+ TN+ FP+ FN)
2	Recall (Sensitivity)	The proportion at which the model predicted "dropout" of the actual correct answer was "dropout" TP/(TP+ FN)
3	Precision	The proportion of the model's predictions of "dropout" that the actual correct answer is dropout". TP/(TP+FP)
4	F-measure	How precise your classifier is, as well as how robust it is. 2 * (Precision * Recall) / (Precision + Recall)
5	ROC curve	The TP ratio and FP ratio graphs are called ROCs. The area under the ROC is called AUC, and the closer the AUC is to 1, the better performing model.

The above evaluation method is based on the Confusion matrix as shown in Table 3, and each indicator used in the evaluation is as follows:

- True Positive (TP): dropout (actual) → dropout (prediction)
- False Positive (FP): non-dropout (actual) → dropout (prediction) Misjudged
- False Negative (FN): dropout (actual) → non-dropout (prediction) Misjudged
- True Negative (TN): non-dropout (actual) → non-dropout (prediction)

TABLE III
INDIVATOR OF CONFUSION MATRIX

		Predictive Value	
		Positive(1)	Negative(0)
Actual Value	True(1)	True Positive(Tp)	False Negative(FN)
	False(0)	False Positive(FP)	True Negative(TN)

F. Design Deep Neural Network

The development of the model used in this work was made using the Keras and Tensor Flow libraries, making it simple to implement except for deep learning models. However, for deep learning models, steps to design deep learning networks are essential and performance differences arise depending on their design methods. The resulting deep learning deep neural network model used in this work is shown in Fig. 4.

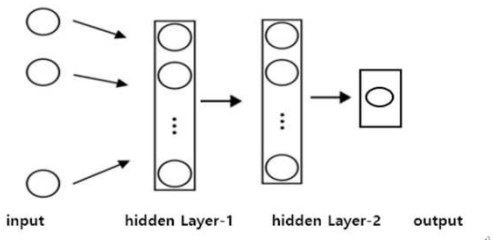

Fig. 4 Design of Deep Neural Network

In DNN (Deep Neural Network) models, the cost function was used as binary_crossentropy, 22 input parameters, 3 hidden layers, and the output function between layers was used as Relu. The last output layer used the sigmodi function. Furthermore, the optimization function chose Adam.

III. Results and Discussion

The experiments collected and used 1,390,650 actual log records and historical data from learning data of 98,685(98,685 * 15weeks) students from March 2012 to December 2019 as follow Fig. 5.

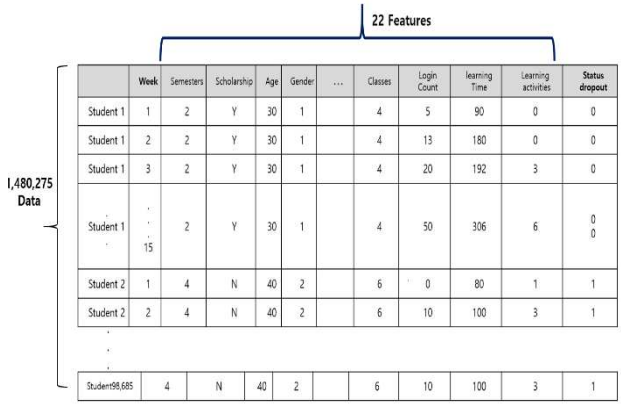

Fig. 5 Sample of data to be used in this study

Experiments performed 7:3 separation of collected and pretreated data into the learning dataset and test datasets, which led to learning on decision tree models, random forest models, support vector machine models, and DNN models.

TABLE IV
Data Set Used Experiment

Data Set	Dropout (1)	Non-Dropout (0)	Total
Train	166,855	869,337	1,036,192
Test	71,509	372,573	444,083
Sum	238,364	1,241,910	1,480,275

We verify the performance of the prediction by applying a tester to the learned model. The performance verification results for each model are as follows:

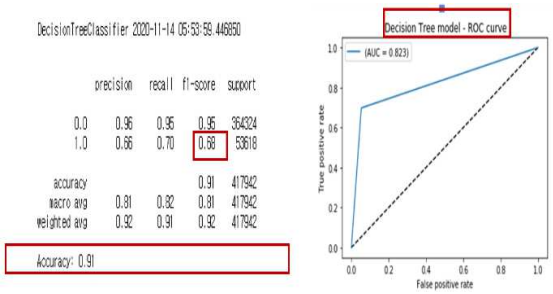

Fig. 6 Decision Tree Model

The results of the prediction verification using the Decision Tree Model show significantly higher performance with 0.91% accuracy and AUC of 0.82 as shown above. However, the slow learning rate is expected to cause processing time and memory problems when applied to actual prediction systems.

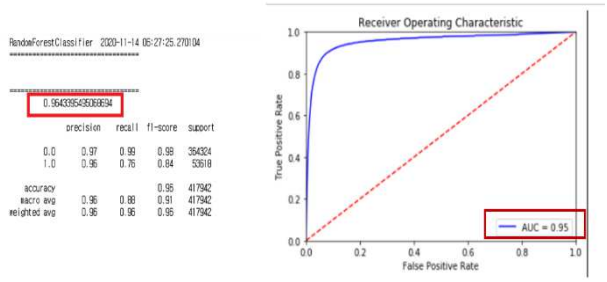

Fig. 7 Random-Forest Model

The following Random-Forest models showed the best performance with accuracy of 0.96%, AUC of 0.95 and the fastest processing speed.

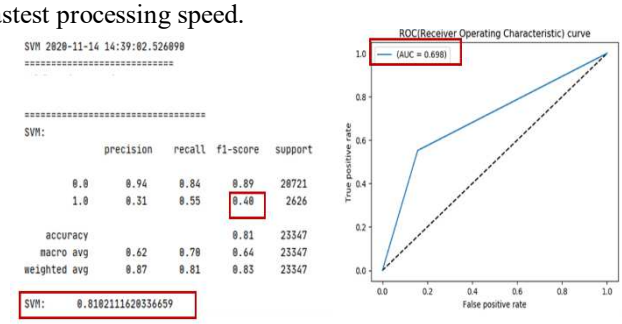

Fig. 8 Support Vector Machine Model

The following Support Vector Machine models showed the lowest performance with 0.81% accuracy and AUC 0.69 performance. It also took too much time to learn, which was not appropriate for the university's prediction system.

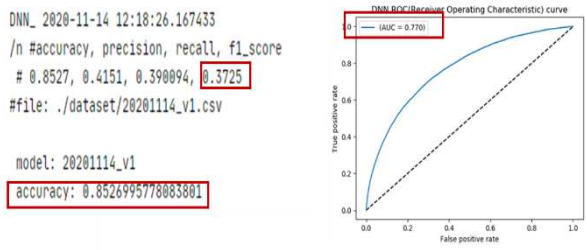

Fig. 9 DNN Model

Along with developing predictive models of good performance, this work aims to identify the potential for the application of deep learning methodologies. Experiments show that we used the most basic deep learning algorithm, but with an accuracy of 0.85% and AUC 0.77, we were able to demonstrate excellent performance.

The results of the experiment are as shown in the table 5 below.

TABLE V
Result of Evaluation

Evaluation	Decision Tree	Random Forest	SVM	DNN
Accuracy	0.91	0.96	0.81	0.85
Recall	0.70	0.76	0.55	0.39
Precision	0.66	0.96	0.31	0.41
F-measure	0.68	0.84	0.40	0.37
AUC	0.82	0.95	0.69	0.77

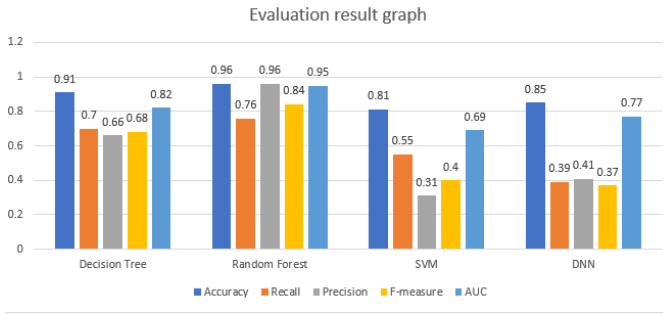

Fig. 10 Evaluation Result graph

As shown in this table, random forest models perform the best with accuracy of 96% , F-measure 0.84% and ACU 0.95%, while deep learning models (DNN) exhibit accuracy of 0.85% and ACU 0.77 values with Adam and three hidden layers of optimization features.

IV. Conclusion

This study developed and validated machine learning and deep learning models to predict early dropout of online learners. In particular, we developed a model that predicts early dropouts by checking the performance every week, so that weekly predictions can be performed to prevent early dropouts in online learning, and in this case, it was confirmed that Random-Forest is the most effective.

In addition, we investigated the possibility of utilizing early dropout predictions of DNN models through various experiments through hyper-parameters tuning of deep learning. As a result, a high accuracy of 85% was confirmed, and with further performance gains, it is expected that deep learning techniques can be used to predict early dropout of online learners. We plan to improve the applicability and performance of various deep learning models for distributed prediction using CNN and RNN among deep learning techniques in the future.

References

[1] Eun-mo Sung, Sung-Hee Jin, and Mi-na Yoo, “Exploring Learning Data for Supporting Self-Directed Learning in the Perspective of Learning Analytics,” *Journal of Educational Technology*, vol. 32, no. 3, pp. 487-533, Sep. 2016.

[2] Ya-Han Hu, Chia-Lun Lo, Sheng-Pao Shih, "Developing early warning systems to predict students’ online learning performance," *Computers in Human Behavior*, vol.36, pp. 469-478, 2014.

[3] Jae-Hoon Han, Suk-Jin Kwon, Jong-Sun Park," (Re)Binding the Factors Affecting Student Learning Outcomes in a Cyber University Using the 3P Model: Learning Analytics Approaches," *Korean Association for Educational Information and Media*, vol. 21, no. 2, pp. 309-332, Jun. 2015.

[4] Jae-won Choi, “A study on the Use of Data Science in Learning Process Management: Focusing on the Risk Prediction,” *(Masters dissertation). Ajou University, Seoul, Korea*, 2016.

[5] W. Xing, D. Du, “Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention,” *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547-570, 2019.

[6] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, N. Radi, "Machine learning approaches to predict learning outcomes in Massive open online courses," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2017-May, no. 7965922, pp. 713-720, Jun 2017.

[7] Y. Lee, D. Shin, H. Loh, (...), B. Kim, Y. Choi, "Deep attentive study session dropout prediction in mobile learning environment," *CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education*, vo. 1, pp. 26-35, May 2020.

[8] B. E. Shelton, J.-L. Hung, P.R. Lowenthal, "Predicting student success by modeling student interaction in asynchronous online courses," *Distance Education*, vol. 38, pp. 59-69, Jan 2017.

[9] F. Dalipi, A.S. Imran, Z.Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," *IEEE Global Engineering Education Conference, EDUCON,* vol. 2018-April, pp. 1007-1014, May 2018.

[10] J. Gardner, C. Brooks, “Student success prediction in MOOCs,” User Modeling and User-Adapted Interaction, vol. 28, pp. 127-203, Jun 2018.

[11] O. W. Adejo, T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," Journal of Applied Research in Higher Education, vol. 10, pp. 61-75, 2018.

[12] V. L. Miguéis, A. Freitas, P. J. V. Garcia, A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decision Support Systems*, vol. 115, pp. 36-51, Nov 2018.

[13] Y.Qu, B. Fang, W. Zhang, R. Tang, M. Niu, H. Guo, Y. Yu, X. He, "Product-based neural networks for user response prediction over multi-field categorical data," *ACM Transactions on Information Systems*, vol. 37, art. no. a3, Oct 2018.

[14] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores Open Access," *Computational Intelligence and Neuroscience*, vol. 2018, art. no. 6347186, 2018.

[15] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences (Switzerland),* vol. 10, art. no. 1042, Feb 2020.

[16] J. A. Ruipérez-Valiente, R. Cobos, P. J. Muñoz-Merino, Á. Andujar, C.D. Kloos, "Early prediction and variable importance of certificate accomplishment in a MOOC," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10254 LNCS, pp. 263-272, May 2017.

[17] W. Zhang, X. Huang, S. Wang, J. Shu, H. Liu, H. Chen, "Student performance prediction via online learning behavior analytics," *Proceedings - 2017 International Symposium on Educational Technology*, art. no. 8005410, pp. 153-157, Aug 2017.

[18] J.-L. Hung, B. E. Shelton, J. Yang, X. Du, "Improving Predictive Modeling for At-Risk Student Identification: A Multistage Approach," *IEEE Transactions on Learning Technologies*, vol.12, no.2, art. no. 8691494, pp. 148-157, 2019.

[19] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, C. Stewart, "Predicting MOOCs dropout using only two easily obtainable features from the first week’s activities," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11528 LNCS, pp. 163-173, 2019.

[20] D. Koehn, S. Lessmann, M. Schaal, "Predicting online shopping behaviour from clickstream data using deep learning," *Expert Systems with Applications*, vol. 150, art. no. 113342, Jul 2020.

[21] X. Ma, Z. Zhou, "Student pass rates prediction using optimized support vector machine and decision tree," *2018 IEEE 8th Annual Computing and Communication Workshop and Conference*, vol. 2018-January, pp. 209-215, Jan 2018.

[22] C. -H. Yu, J. Wu, A. -C. Liu, "Predicting learning outcomes with MOOC clickstreams," *Education Sciences*, vol. 9, art. no. 104, Jun 2019.

[23] K. Limsathitwong, K. Tiwatthanont, T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science*,pp. 110-114, May 2018.

[24] G. Kostopoulos, S. Kotsiantis, C. Pierrakeas, G. Koutsonikos, G. A. Gravvanis, "Forecasting students' success in an open university," *International Journal of Learning Technology*, vol. 13, pp.26-43, 2018.

[25] L. C. Sorensen, ""Big Data" in Educational Administration: An Application for Predicting School Dropout Risk," *Educational Administration Quarterly*, vol. 55, pp. 404-446, 2019.

[26] K. Coussement, M. Phan, A. De Caigny, D. FBenoit, A. Raes, "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model," *Decision Support Systems*, vol.135, art. no. 113325, 2020.

[27] N. R. Aljohani, A. Fayoumi, S. -U. Hassan, "Predicting at-risk students using clickstream data in the virtual learning environment," *Sustainability (Switzerland),* vol.11, art. no. 7238, 2019.

[28] R. Raga, J. Raga, "Early prediction of student performance in blended learning courses using deep neural networks," *Proceedings - 2019 International Symposium on Educational Technology, ISET 2019*, art. no. 8782240, pp. 39-43, 2019.

[29] J. Lagus, K. Longi, A. Klami, A. Hellas,, "Transfer-learning methods in programming course outcome prediction," *ACM Transactions on Computing Education*, vol. 18, art. no. 19, 2018.

[30] P. E. Ramírez, E. E. Grandón, "Prediction of student dropout in a Chilean public university through classification based on decision trees with optimized parameters," *Formacion Universitaria*, vol. 11, pp. 3-10, Jun 2018