IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2021, accepted January 12, 2022, date of publication January 18, 2022, date of current version January 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144625

Online Students' Learning Behaviors and Academic Success: An Analysis of LMS Log Data From Flipped Classrooms via Regularization

JIN EUN YOO[1], MINJEONG RHO[1], AND YEKYUNG LEE[2]
[1]Department of Education, Korea National University of Education, Cheongju 28173, Republic of Korea
[2]Graduate School of Education, Sogang University, Seoul 04107, Republic of Korea

Corresponding author: Minjeong Rho (minjeong019@gmail.com)

This work involved human subjects in its research. Approval of all ethical procedures and protocols was granted by the Korea National University of Education Institutional Review Board under Application No. KNUE-2019-H-00093.

ABSTRACT The main purpose of this study was to demonstrate the uses of regularization, a machine learning technique, in exploring important predictors for online student success. We analyzed student and learning behavioral variables from undergraduate fully-online flipped classrooms. In particular, students' instructional video watching behaviors at an instructional unit level were extracted from LMS (learning management system) log data, and Enet (elastic net) and Mnet were employed among regularization. As results, regularization not only showed comparable prediction performance to random forest, a nonlinear method well-known for its prediction capabilities, but also produced interpretable prediction models as a linear method. Enet and Mnet selected 17 and 19 important predictors out of 159, respectively, and could identify potential low-performers as early as the first instructional week of the course. Important variables rarely recognized in previous studies included complete viewings of the first video before class and repeated complete viewings of challenging contents after in-class meetings. Unlike previous studies, aggregate measures of video lecture views were not important predictors. Variables less frequently studies in previous studies were the number of non-mandatory quiz-taking and mobile lecture watching frequencies. Variables in line with previous research were student attitudes towards the course, gender, grade level, and clicks on learning materials postings. Many students turned out not to watch lecture videos completely before class. Further research on regularization and exploration of these variables with other potentially important predictors can provide more insight into students' online learning from a comprehensive perspective.

INDEX TERMS LMS log data, machine learning, regularization, random forest, flipped classroom, online learning, learning analytics.

I. INTRODUCTION

The COVID-19 pandemic has changed the education system worldwide. Online learning is no longer an option, and an increasing number of online classes have incorporated components of flipped classrooms (FC) in an effort to improve the quality of learning and instruction. Despite varying results regarding the effectiveness of flipped learning in higher education [1]–[4], FC has grown rapidly as an innovative pedagogical approach in recent decades. FCs are designed to integrate in-class activities (e.g., group discussions) and out-of-class activities (e.g., watching lecture videos in advance) to accomplish high levels of academic achievement. Thus, in FCs, students' self-directed pre-class activities are greatly emphasized as a necessary condition to enhance in-class learning and instruction [5].

However, there has been little empirical research using quantitative observational data [6] for specifically identifying FC learning behaviors significant for academic success. This may relate to methodological limitations of previous research analyzing LMS (learning management systems) log data in terms of data collection and analysis methods. First of all, particularly in traditional FC designs LMS cannot capture all the students' learning activities; data from pre-class

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

assignments representing prior engagement in learning typically exist outside LMS. However in fully-online FCs, which are increasingly prevalent in the COVID-19 situation, collecting trace data has become much easier, as students' pre-class video watching activities, for instance, are stored in LMS. If LMS log data representing learners' study behaviors and patterns are analyzed, researchers could discover unknown relationships among many variables explaining learning in online environments.

Furthermore, it is possible to unobtrusively collect near-real-time information through LMS; students' behaviors in LMS are automatically stored in the log files without the students' cognizance [7]–[10]. Previous research collected data from ex post facto self-report surveys asking students how well pre-class assignments were carried out (e.g., [11]), which is meaningful to some extent. However, self-report questionnaires rely on memories and reflections, and thus are prone to social desirability bias.

Second, analysis methods have room for improvement; LMS log data of students' learning behaviors have not been utilized to its full potential. Despite the aforementioned advantages that log data bring to data analyses, the intractability of LMS log data has been a practical hindrance. Log data are unstructured, which can lead to high-dimensional data (i.e., more variables than observations), depending on data pre-processing and cleaning. Relatedly, previous studies employing traditional methods (e.g., [12], [13]) or early ML techniques (e.g., [14], [15]) have analyzed aggregate variables such as total login frequencies or average login hours, which contributed to preventing possible problems of traditional methods combined with high-dimensional data such as convergence. However, instructional unit-based data traced from log data can serve as better indicators of study behaviors. For instance, study habits of online students with high levels of academic success can be observed even in the first few weeks of a course [16]. By implementing instructional unit-based analysis, more can be learned about when and how instructor intervention should be provided during the semester.

When behavioral variables at an instructional level are to be analyzed, more advanced ML appears to be a necessary technique due to its capacity to handle possible high-dimensional data without convergence problems. In the similar vein, a large number of predictors can be explored in one ML model, which in turn propagates the creation of a new theory or complements existing ones [17]. Extensive modeling with as many predictors as possible via ML appears necessary to explore yet uninvestigated important variables to predict students' academic achievement. Of note, compared to the traditional OLS (ordinary least squares) regression, nonlinear ML methods such as random forest, support vector machines, and deep learning consist of complex higher-order interactions, and do not provide explainable prediction models. Nonetheless, learning analytics is one of the fields which needs to be augmented with explanation.

We propose regularization (penalized regression) among ML to analyze LMS log data. There has been little research employing regularization methods in LMS log data analysis, but regularization can contribute to the field in that it produces interpretable prediction models [18]–[20]. Based on linear regression, the regression coefficients of regularization can be interpreted similarly to those in traditional, non-penalized regression. This is a great advantage in learning analytics, as prediction models need to be interpreted under certain educational settings, for instance to plan more effective intervention strategies for at-risk students. While regularization produces interpretable prediction models, nonlinear ML methods may outperform regularization in terms of prediction. Therefore, it was worthwhile to compare regularization to random forest, a popular nonlinear ML method famous for its prediction capabilities.

In summary, this study examined the prediction performance of ML methods for online student success, and explored important learning behaviors in fully online flipped classrooms at an instructional unit level. Specifically, the following research questions were posed.

1. Which ML technique, random forest or regularization, shows better performance in analyzing LMS log data in terms of prediction?

2. In fully-online flipped classrooms, which learning behaviors extracted from LMS are important for predicting students' academic achievement?

3. In fully-online flipped classrooms, what are the students' video watching patterns at an instructional unit level?

Although the focus of this study was on video watching behaviors at an instructional unit level, we endeavored to include as many variables as possible in predictive modeling 1) to fully utilize the strengths of ML and 2) to compare the importance of video watching behaviors to that of other variables. In particular, students' gender, grade level, and attitudes toward the course were investigated as well as class material downloads, forum postings, and quiz-taking via PC or mobile. Details of the variables in this study are explained in IV-C.

II. LITERATURE REVIEW

A. RECENT ML RESEARCH IN LEARNING ANALYTICS

An increasing number of studies have started to employ ML in predicting student success. Although regularization has been in popular use as an approach to predictive modeling in diverse fields including bioinformatics (e.g., [21], [22]) and engineering (e.g., [23], [24]), less application of penalized regression to learning analytics data appears in the literature. Bertolini [25] summarized a total of 10 recent learning analytics studies. All the 10 studies employed nonparametric ML methods such as SVM (support vector machines), ANN (artificial neural networks), tree, RF (random forest), and gradient boosting; but only a few studies compared the nonparametric methods to regularization. When employed, however, regularization showed comparable or better prediction

performance. For instance, in a study by Beemer *et al.*[18] lasso (regularization) showed better prediction performance than RF in predicting binary variables, and was next to RF in predicting continuous variables.

Other studies in learning analytics have utilized ML in the framework of traditional methods. For instance, Wu [26] labeled students' text data with ML, and the labeled variable was used as a predictor in traditional regression to explain 78 students' academic performance in a statistics course. Specifically, he employed weakly supervised ML to rate students' Facebook posts and comments. In order to categorize the students' text data into four ordinal groups of relevance, an ensemble ML classifier was created, consisting of RF, ANN, and SVM. This ML rating served as one of the nine independent variables in different OLS regression models. The model containing the ML rating variable outperformed that of human coding variable in terms of model fitting criteria such as R-squares and BIC. The ML rating also showed the highest effect size, followed by help-seeking tendency, and procrastination. On the other hand, students' internet and Facebook time, which were aggregate variables obtained from students' self-report, were not statistically significant.

Bosch[27] explored 51 predictors of over 10,000 students from a quasi-experimental study with gradient boosting (a nonlinear ML). He inspected the top 4 to 5 predictors of highest importance from gradient boosting, but there is no solid rule on the cut-off. He also calculated Pearson's r values and reported statistical significance. This hybrid approach of ML and significance testing is understandable in that nonparametric ML models lack interpretability. However, ML and conventional methods stem from different standpoints. While conventional methods value explanation, prediction is the goal of ML. The important predictors extracted from ML best serve the purpose of prediction, not explanation. Therefore, the selected top 4 to 5 predictors were unlikely to be in the order of the highest statistical significance (p. 6). Bosch also reported that OLS regression showed comparable performance to gradient boosting; nonlinear ML may be overqualified for experimental data obtained in a traditional framework.

In summary, many predictive modeling studies in learning analytics have focused on the comparison of nonlinear ML methods such as SVM, ANN, tree, RF, and gradient boosting. However in few studies which included regularization in the comparison set, nonlinear ML did not outperform regularization in terms of prediction performance. In other recent studies of learning analytics, ML has been utilized coupled with traditional methods including OLS regression [26] or Pearson correlation [27], after serving the purpose of feature engineering [26] or variable selection [27].

B. PREDICTING EFFECTIVE LEARNING BEHAVIORS IN FLIPPED CLASSROOMS

In a flipped classroom (FC), students carry out self-study outside the classroom and then engage in interactive learning activities during their in-class meetings. The most critical aspect of FC is that it is systematically designed to engage learners in self-regulated learning out of the class that culminates into higher learning achievements. For instance, they must exert self-directed effort into pre-class learning, engage in online activities such as posting their ideas, taking quizzes, and reviewing class materials so that in-class time is not wasted. These activities require self-regulated learning (SRL), which implies the learner's active engagement from a metacognitive, motivational, and behavioral point of view [28].

Since completing pre-class assignments and preparing for interactive in-class activities are critical in FC, a high level of SRL is necessary for students to succeed. SRL strategies such as effective time management, metacognition, and effort regulation are considered significant predictors of academic success [29]. Pintrich [30] proposed a conceptual framework for SRL composed of four phases: 1) Forethought, planning and activation, 2) Monitoring, 3) Control, and 4) Reaction and reflection. Behaviors representing each phase are time/effort planning, self-observation, increase/decrease of effort, and persistence [30]. In terms of pre-class learning behaviors in FCs, students must plan ahead their time for watching lecture videos, watch lectures, monitor their understanding of the contents, go over the lecture again until understanding is complete.

Previous studies examining student engagement in learning have used variables reflecting individual LMS usage, but not necessarily SRL. SRL is based on the reciprocity between the learner and the context of learning. Thus, variables for SRL should use data representing the processes of learning rather than aggregate data representing total usage. LMS usage data are usually collected by aggregate measures of login frequencies, menu usage, material download, content pages viewed, and posted messages [10], [12]–[15], [31], [32]. Aggregate measures of these data have displayed inconsistent effects on student achievement. Login frequencies [13], [15] and LMS menu usage [12]–[14] were statistically significant or important indicators of students' academic achievement in online learning. In contrast, in MOOC (massive open online course) environments, forum variables such as numbers of messages posted, or comments received were not directly related to students' learning [32]. Students' instructional video watching behaviors derived from LMS have been another important measure of SRL in MOOC [31], [32] or FC settings [33]. Studies on MOOCs obtained video watching behavior data from access to videos [31], percentages of opened and completed videos [32], and percentages of video play (or pause) actions within a study session [33]. Of note, these variables on video watching were also measured in the form of aggregate variables such as percentages of video viewed or played [32], [33] and total user access to videos [31]. These variables did not significantly improve the predictive power of the models, particularly when exam (exercise) variables were present [32].

Although aggregate data for pre-class learning time are useful for examining student engagement, more research is necessary to understand the patterns of students' learning behaviors for successful FCs. Specifically, variables associated with SRL should manifest the behaviors related to the four phases of SRL. For instance, watching video lectures completely several times after a difficult class may be a sign of self-observation, followed by increased effort for mastering difficult contents. Thus, using instructional unit variables rather than aggregate variables may be a solution for measuring SRL behaviors from LMS log data. Students' behavioral data at an instructional unit level reveal patterns/processes of their studies across the phases of the semester, contents, exams, and such. This type of data can be indicative of SRL and predict academic achievement with higher accuracy.

Studies analyzing instructional unit variables or the like have demonstrated patterns of SRL behaviors in online settings. Considering time management a significant SRL factor for predicting academic performance, Cerezo *et al.* [34] showed that early access to the first theoretical resource, medium amount of time on assignments, and access to assignments within average time-frame lead to high performance. In other words, examining behavioral variables in relevance to changes in the instructional context throughout the semester can provide new insights into learning in online environments. Jovanovic *et al.* [35] suggested that course-specific indicators could be better predictors of academic success by comparing the predictive power of generic versus course-specific indicators, and demonstrated that the latter, specifically regularity of pre-class activities, had higher predictive power. To conclude, the highest meaning of learning analytics can be achieved when learning analytics is executed within the instructional context. Analyzing instructional unit data together with course-specific contexts can provide us more insight into effective learning behaviors.

III. MACHINE LEARNING FOR PREDICTION MODELING

Among ML, we employed regularization and RF. While regularization produces linear, parametric models, RF is categorized as a nonlinear, nonparametric method. Linear models are easier to interpret than nonlinear models, but nonlinear models may have strengths in prediction.

A. REGULARIZATION: ENET AND MNET

The purpose of regularization is to reduce the mean squared error (MSE) and the subsequent prediction error by introducing a small bias, thereby lowering variance in the estimates [36]. Regularization imposes a penalty function on top of the objective function, and shrinks some coefficients to zero. We chose elastic net (Enet) and Mnet for the following reasons. First, Enet is a combination of LASSO and ridge [37], and Mnet is a combination of MCP and ridge [38]. By incorporating ridge in the models, both handle multicollinearity, a likely challenge in LMS log data analysis. Second, Enet and Mnet represent convex and concave regularization, respectively. While Enet has the issue of inconsistency, Mnet is known to produce nearly consistent estimates [38].

Consider a linear regression model with p predictors and n observations. Suppose the predictors are divided into K non-overlapping groups, the response variable y is an n-dimensional vector, and X_k is an $n \times p_k$ design matrix of the p_k predictors in the k-th group. β_k is the p_k-dimensional vector of regression coefficients of the k-th group, and ϵ is an n-dimensional vector of mean zero (equation (1)).

$$y = \sum_{k=1}^{K} X_k \beta_k + \epsilon. \tag{1}$$

For a Gaussian family, Enet and Mnet are expressed as in equations (2) and (3), respectively. The same first term on the right-hand side of equations (2) and (3) is the loss function of least squares. The second term on the right-hand side of equation (2) is the penalty function of Enet, consisting of two tuning parameters: λ and α. The parameter λ regularizes shrinkage of the coefficients, and the parameter α controls the amount of ridge. Typically, α is set to be 0.5 for collinear data [39].

The second and the last terms of equation (3) are the MCP and ridge penalties, respectively. The parameter λ_1 controls the amount of the penalty. The parameter γ, known as the concavity penalty, regulates the penalization rate depending on the size of the coefficients. When the coefficients are larger than the product of the two penalties, the rate of the MCP penalty quickly drops, thereby applying less shrinkage to the coefficients and yielding less biased estimates than LASSO does [19], [38]. Lastly, the ridge penalty is added to the Mnet equation, the parameter of which is λ_2.

$$\hat{\beta}^{Enet} = argmin_\beta \left[\frac{1}{2n} ||y - \sum_{k=1}^{K} X_k \beta_k||^2 + \lambda \sum_{k=1}^{K} (\alpha ||\beta_k||_1 + (1-\alpha)||\beta_k||_2^2) \right].$$

$$\hat{\beta}^{Mnet} = argmin_\beta \left[\frac{1}{2n} ||y - \sum_{k=1}^{K} X_k \beta_k||^2 + \sum_{k=1}^{K} J(||\beta_k||_1 | \lambda_1, \gamma) + \lambda_2 \sum_{k=1}^{K} ||\beta_k||_2^2 \right], \tag{2}$$

where

$$J(x|\lambda_1, \gamma) = \begin{cases} -\frac{1}{2\gamma} x^2 + \lambda_1 |x|, & |x| \le \gamma \lambda_1, \\ \frac{1}{2} \gamma \lambda_1^2, & |x| > \gamma \lambda_1. \end{cases} \tag{3}$$

1) CROSS-VALIDATION (CV)

In particular, this study executed subsampling techniques for variable selection in order to consider the bias resulting from data-splitting in model validation [40], [41]. The steps were as follows. First, the whole data were randomly divided with the ratio of 7:3 to get the training and test data, respectively.

This resulted in high-dimensional test data (71 students of 159 predictors), while training data consisted of 166 students of 159 predictors. Data characteristics are explained in IV-C2. Second, 5-fold CV (cross-validation) was executed on the training data. For a value of the penalty parameter, the training data were split with the ratio of 4:1. The 4/5 or 4 folds of the training data were used in model fitting and the 1/5 or remaining fold was used in model evaluation, which is repeated for every fold. The prediction error of the λ is calculated, which is referred to as the CV error of the λ (equation (4)) [42]. Third, the second step was repeated for every λ in the range, and the λ of the lowest CV error served as the penalty value of regularization. That λ value was applied to the test data in step 1, which yielded prediction errors.

$$CV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i^{(\lambda)})^2. \quad (4)$$

2) SELECTION COUNTS

Of note, the variables of nonzero coefficients after regularization should not be interpreted as 'statistically significant.' Regularization produces biased estimates, and significance testing is performed on unbiased estimates. Special techniques such as post-selection inference (e.g., [43]) are required to perform statistical testing after regularization, but currently only available with LASSO [19], [20]. Instead of statistical testing, we iterated data splitting and prediction modeling, and obtained selection counts as the criterion for variable importance; variables selected more often bear more importance than variables selected less often [19], [20]. The aforementioned three steps were repeated 1,000 times with random data-splitting. The selection or non-selection of each variable from the second step was counted in the 1,000 iterations, which served as the selection counts of the study. Specifically, this study presented variables selected 1, 250, 500, 750 times or more, and all 1,000 times. Yoo and Rho[20] suggest that variables selected 75% or more with Enet and 50% or more with Mnet are worth investigation in a simulation study of social science large-scale data, but there is no such study in the context of learning analytics. All the programs were written in R 3.6.2. Specifically, the grpreg library [44] was used for regularization.

B. RANDOM FOREST AND FINE TUNING

The first research question of this study was to compare the prediction performance of regularization to that of random forest (RF). RF models are known to be highly predictive but difficult to interpret. With base learner as decision trees, RF yields nonparametric models. RF creates bootstrapped samples, fits decision tree models on the bootstrapped samples, and combines the decision tree results as an ensemble method [45]. Tuning parameters of RF include the number of variables randomly sampled as candidates at each split (mtry), the number of trees, the minimum number of observations in a terminal node, sampling with or without replacement, and splitting criteria [46]. Among them, mtry is reported to affect the complexity of the final model [47]. In particular, small mtry decorrelates trees, which leads to models of low variance and high bias [46], [47]. Thus, models from small mtry tend to be stable, but important predictors might be excluded in modeling, resulting in decreased prediction [46]. Despite its importance, however, there has not been enough empirical research on the proper values of mtry; researchers typically adopt the default values that Breiman [45], the inventor of tree and RF, suggested.

In an effort to maximize the prediction performance of RF, we tuned mtry in all possible range, using OOB (out-of-bag) errors as the evaluation criteria. We set the number of trees to 5,000. This is ten times of the default in the random Forest library in R [48], and is considered sufficiently large to yield stable prediction results given the sample size. The other tuning parameters adopted the default values of the random Forest library. The steps were as follows. First, the same sets of training and test data as in regularization (III-A1) were utilized in model fitting. Second, the mtry parameter was tuned in the range of 1 (a stump model) and 159 (all the predictors in the model). For an mtry value in the range, the training data were fitted and the mtry value of the lowest OOB errors were identified. Third, the mtry value from step 2 was applied to the test data from step 1, and the RMSE (root mean square error) of the test data was calculated, which served as the prediction error of the RF model. As was with regularization, all three steps were iterated 1,000 times. The OOBCurve package [49] in R was used.

Of note, we also obtained prediction errors using Breiman's default. As there were 159 predictors, the default value of mtry was 53 (=159 divided by 3). The aforementioned three steps were employed, except that the mtry value was set to 53 in step 2; the other steps were the same. The comparison of what Breiman [45] suggested and what CV yielded was expected to give another insight into the tuning process of RF in analyzing LMS log data.

IV. MATERIAL AND METHODS

A. PARTICIPANTS

In the Fall semester of 2020, 242 undergraduate students in a pre-service teacher program enrolled in 8 fully-online undergraduate classes titled Measurement and Evaluation. Students were required to complete 6 out of 9 prerequisite classes to receive their teaching certificate, and Measurement and Evaluation was one of the 9 classes. The classes of the Fall semester were for sophomores majoring in Liberal Arts and Social Sciences, but a small number of off-semester or off-grade students who missed taking the classes on time due to leave of absence or schedule conflict were also allowed to enroll in the classes. The male-to-female ratio was 37.13% (88) to 62.87% (149). On-grade to off-grade student ratio was 85.65% (203) to 14.34% (34). On-semester to off-semester student ratio was 93.25% (221) to 6.75% (16). Students on average had 1.38 times of practicum (SD = 0.61).

B. SETUP OF THE FC

Three instructors (A, B, C) including a head-instructor (A) taught the 8 classes; A taught 4 classes and B and C two classes each. Before the semester started, they held several meetings to discuss details including class progress, team projects, and the final exam. As a result, all the 8 classes scheduled a simultaneous final exam at the end of the course, and shared the same class materials including instructional videos, textbooks, and syllabus. The syllabus clearly stated that the course would apply flipped learning. On the orientation day of the first week, each instructor gave a detailed overview of FC and its potential benefits. The importance of the weekly assignments of instructional video watching before class was also emphasized, particularly because students were asked to create and complete class projects within groups based on the contents of the assigned videos.

The instructional videos were pre-recorded PowerPoint presentations carried out by the head-instructor, with content based on a book also written by the head-instructor. A total of 34 video clips covered 11 instructional weeks, and the numbers of 1 to 11 in the video names indicate the corresponding instructional weeks (refer to the videos 1_1 to 11_4 in Appendix). The mean length of the 34 videos was about 11 minutes with an SD of 5.92. The minimum and maximum values were 4.1 and 29.2, respectively, but most of them ranged between 5 and 10 minutes. The first, second, and third quartiles of the video length were 7.13, 9.11, and 13.65, respectively, indicating a right-skewed distribution. Each week's running time was between 23.81 and 52 minutes, and students on average were expected to watch 34.34 minutes of videos each week (SD = 8.59). The first, second, and third quartiles of the weekly length were 28.84, 32.90, and 38.61, respectively.

During class, interactions in small groups of 4 to 6 students were greatly encouraged. The groups engaged in discussions on team projects and SPSS exercises in Zoom breakout rooms. The instructor observed group interactions and at the end of the class gave short lectures on some of the topics that students appeared to have developed misconceptions about. A non-mandatory quiz of 3-4 short questions was also presented before class for each instructional week in LMS. Students were told that the quizzes would serve as formative assessments and the quiz scores did not count toward grades.

C. DATA

1) LMS LOG DATA

In total, 21,589 rows of video watching activities as well as 5,107 rows on board-posts readings were recorded in the log file. As many of the students indicated that they used the double-speed option of the LMS in video watching, we used 50% of the video length as a criterion. If a student watched a video 50% of the length or more, the student is counted to have completed watching the video, and vice versa.

As the second and third research questions related to students' video watching behaviors at an instructional unit level, this study counted the frequencies of each video watching, separating before/after and attempted/completed watching. Specifically, 4 variables were created for each video: BI (incomplete attempt before class), BC (complete watching before class), AI (incomplete attempt after class), and AC (complete watching after class). Eight aggregate variables were also obtained for comparison purposes to previous research: BI, BC, and B (=BI + BC) for before class counts; AI, AC, and A (=AI + AC) for after class counts; and lastly I for all incomplete watching and C for all complete watching. As a result, students' video watching activities were reorganized as 144 variables: 136 (=34 videos 4 variables) plus 8 aggregate variables.

Ten other variables extracted from the log data included the numbers of clicks on SPSS data materials (spss.post, spss.sum), Q and A (qna.post, qna.sum), and other board postings (board.post, board.sum). These variables were summarized in pairs. Variables named *.sum (e.g., spss.sum) indicate the total clicked numbers, while the other pair (e.g., spss.post) only counted once for multiple clicks. The frequencies by device (e.g., mobile or PC) of quiz-taking and video watching were also obtained. Variables test.P and test.M indicate the frequencies of quiz-taking via PC and mobile, respectively, and lecture.P and lecture.M for instructional video watching via PC and mobile, respectively.

2) STUDENT VARIABLES AND RESPONSE VARIABLE

Student variables were measured from a survey administered in LMS during the second week of the course. They included gender (1=male; 0=female), on-grade (1=sophomore; 0=others), on-semester (1=yes; 0=no), number of practicum, and the mean score of an attitude survey. The survey consisted of 25 Likert-scaled items, and measured students' attitudes toward measurement and evaluation (Cronbach's alpha=.92). Sample items are "It is important to know and use performance level descriptions in student evaluation" and "Establishing alignment between curriculum, assessment, and instruction is necessary to student evaluation."

The response variable of this study was the final test score. The final test consisted of 35 multiple-choice items covering the 11 weeks, and was given simultaneously to all the 242 students at the last week of the course. Five students missed the final, and their data were excluded from further analysis. The mean score of the final was 25.6 (SD = 5.68). The lowest score was 5, and the highest (perfect) score was 35. The first, second, and third quartiles of 22, 27, and 30, respectively, indicating a left-skewed distribution. The student variables were merged to the video watching variables from LMS log data (IV-C1), which resulted in the final dataset of 159 predictors and 1 response variable of 237 students.

TABLE 1. Test data RMSE of regularization and RF.

	Min	Q1	Med	Mean(SD)	Q3	Max
Enet	4.43	5.31	5.60	5.61(0.43)	5.92	6.93
Mnet	4.56	5.39	5.66	5.69(0.43)	5.97	8.56
RF-default	4.26	5.18	5.43	5.43(0.38)	5.68	6.94
RF-tuning	4.32	5.19	5.49	5.48(0.40)	5.74	6.98

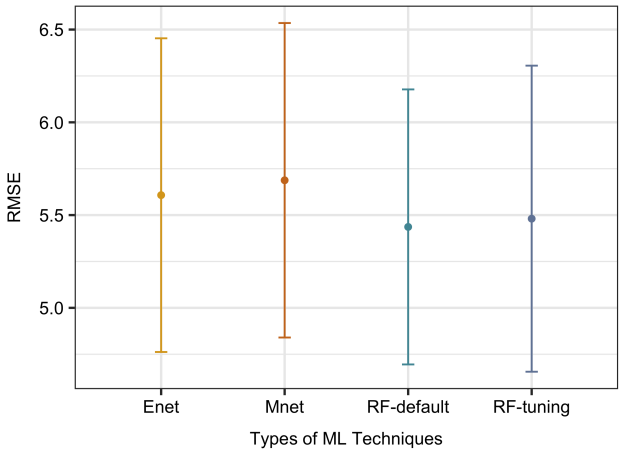

FIGURE 1. 95% Confidence intervals of test data RMSE.

TABLE 2. Descriptive statistics of the mtry values after RF-tuning.

	Min	Q1	Med	Mean(SD)	Q3	Max	Mode
Value	1	17	34	44.67(37.83)	60	158	5, 8

V. RESULTS

A. MACHINE LEARNING RESULTS (RQ1)

1) PREDICTION MEASURE

The first research question was to examine whether RF showed better prediction performance than regularization. Specifically, a total of four ML models, Enet, Mnet, RF-default, and RF-tuning, were compared in terms of test data RMSE. Despite the efforts to maximize prediction by tuning the mtry parameter of RF, the 95% confidence intervals of the four models overlapped (Table 1 and Figure 1). Test data MAE (mean absolute error) showed similar patterns, and the 95% confidence intervals also overlapped.[1] In summary, the four models were not statistically different in terms of prediction.

For better understanding, we present the descriptive statistics of the tuned mtry values across 1,000 iterations in Table 2. The mtry values after tuning covered almost all possible range with the standard deviation (37.83) being close to the mean (44.67) or median (34). They also tended to be smaller than Breiman's suggestion, the default value of 53.

2) SELECTION COUNTS OF REGULARIZATION

As the confidence intervals overlapped (Figure 1) and regularization produces interpretable prediction models, we chose regularization for subsequent analyses. Figure 2 shows the solution paths of Enet and Mnet with a random seed. The horizontal axis depicts λ (penalty parameter) values in range, and the vertical axis indicates the regression coefficients. Each curve in the solution path corresponds to a predictor, and increasing values of λ shrink the coefficients toward zero. While Enet using a convex penalty function imposes the same penalty on the coefficients, the concave penalty of Mnet tapers off with larger coefficients in absolute values, and thus larger coefficients get less shrunken with Mnet than with Enet (see Figure 2).

[1]The MAE results are available upon request.

TABLE 3. Selection counts of regularization.

	≥ 1	≥ 250	≥ 500	≥ 750	=1,000
Enet	135	34	17	1	0
Mnet	106	19	2	0	0

The selection counts of regularization after 1,000 iterations are presented in Table 3. A total of 135 and 106 predictors were selected out of 159 at least once with Enet and Mnet, respectively. Applying 25% or more selection counts resulted in 34 and 19 predictors of Enet and Mnet, respectively. A total of 17 and 2 predictors were selected at least 1 out of 2 runs of Enet and Mnet, respectively. No predictor was selected in all 1,000 iterations with either Enet or Mnet. In sum, Mnet produced larger coefficients and selected fewer variables than Enet, consistent with literature (e.g., [38], [50]).

B. IMPORTANT PREDICTORS OF STUDENT ACHIEVEMENT (RQ2)

The second research question was to investigate important predictors of student achievement. Mnet always selected fewer predictors than Enet, but both Mnet and Enet showed similar trends in selecting important predictors. The summary of predictors selected 50% or more for Enet and 25% or more for Mnet are presented in Table 4. The first three predictors (gender, on-grade, and attitudes) were from a student survey administered in LMS, predictors 8 to 19 were video-watching variables at an instructional unit level, and predictors 4 to 7 were other learning behaviors extracted from LMS log data.

1) VIDEO-WATCHING VARIABLES

Among the 144 variables on video watching, 10 to 12 variables were selected as important depending on the regularization methods. First of all, the very first video (1_1) turned out to convey crucial information in predicting students' achievement. Although it covered the easiest contents, formative assessment (refer to Appendix), the more the students completed watching the first video before class (BC01_1), the higher their final scores were. Specifically, one more completed watching of the first video before class was associated with 0.59 (Enet) and 0.69 (Mnet) higher scores in the final exam. By contrast, one more completed watching of the first video after class (AC01_1) was associated with 0.47 and 0.59 lower scores in Enet and Mnet, respectively.

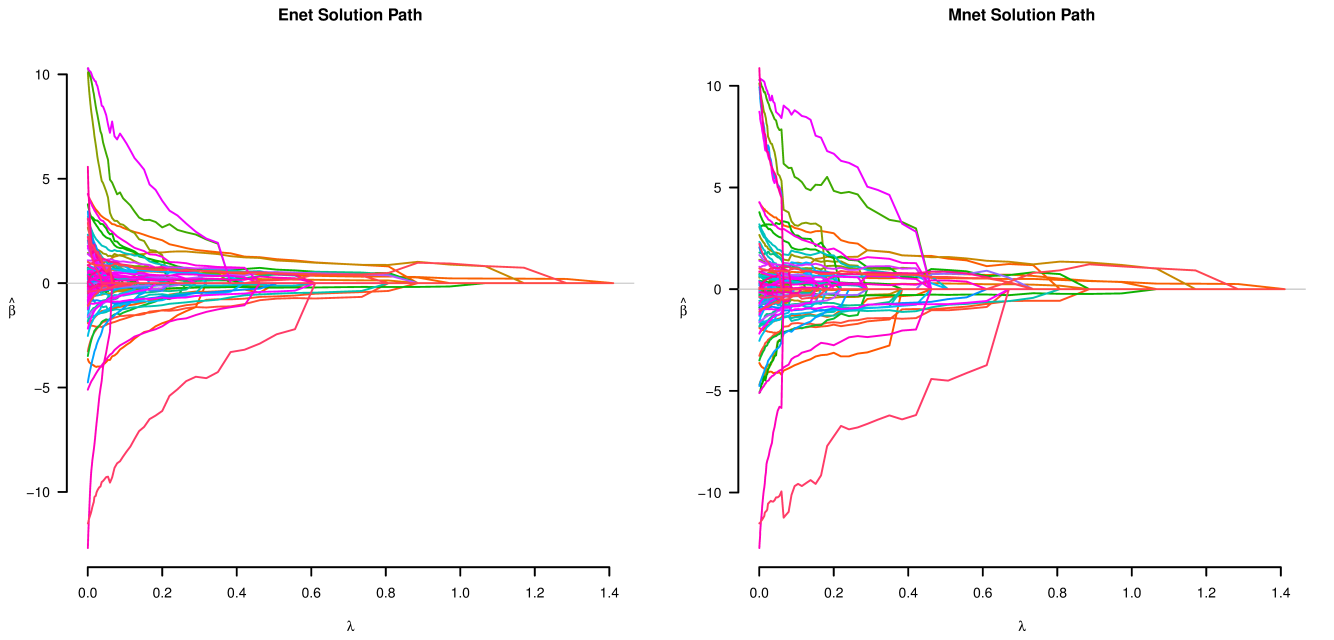

FIGURE 2. Solution paths of Enet and Mnet (seed = 1234).

TABLE 4. Coefficients of selected predictors by regularization.

	variable	Enet			Mnet		
		Mean	SD	#	Mean	SD	#
1	gender(male)	-0.56	0.31	598	-0.70	0.38	341
2	on-grade	-0.58	0.26	731	-0.76	0.32	486
3	attitudes	1.39	0.53	819	1.67	0.63	634
4	test.P	0.23	0.10	691	0.30	0.15	460
5	test.M	0.34	0.12	525	0.50	0.15	306
6	lecture.M	0.01	0.01	516	0.02	0.01	283
7	spss.sum	0.11	0.05	515	0.16	0.07	269
8	BC01_1	0.59	0.29	735	0.69	0.37	497
9	AC01_1	-0.47	0.22	622	-0.59	0.29	366
10	BI02_2	-0.45	0.17	518	-0.63	0.23	278
11	AI04_2	-0.19	0.08	614	-0.25	0.11	369
12	AC04_3	0.38	0.24	503	0.55	0.33	295
13	AC06_2	0.33	0.21	599	0.40	0.29	367
14	AC09_3	0.33	0.20	673	0.45	0.29	436
15	AC10_2	0.43	0.26	704	0.55	0.37	473
16	AC11_4	0.33	0.23	563	0.44	0.34	322
17	AI11_4	0.35	0.17	564	0.45	0.25	307
18	AI04_3				-0.21	0.11	293
19	AC06_1				0.43	0.31	269

Note: # indicates selection counts in 1,000 iterations.

This variable, AC01_1, was the only 'AC' variable having a negative relation to the final exam.

Second, a total of seven 'AC' variables were selected important. With the aforementioned exception of AC01_1, the other AC variables (e.g., AC04_3, AC06_1, AC06_2, AC09_3, AC10_2, AC11_4) had positive relations to the final. These AC variables covered either the earlier unfamiliar, technical contents or the most difficult concepts at the end of the course. Particularly, the unfamiliar, technical contents included Ebel and Angoff standard setting (AC04_3) and the first SPSS practice (AC06_2). Videos in the last weeks of the course (weeks 9 to 11) covered the most difficult concepts such as Cronbach's alpha (AC09_3), reliability with SPSS (AC10_2), and the relationships between reliability and validity (AC11_4). Students who completed watching these earlier technical or later difficult videos multiple times after class were more likely to obtain higher final scores.

Third, two 'B' variables were selected important: BC01_1 and BI02_2. Aforementioned, completing the first video before class (BC01_1) related to higher final scores, but conversely incomplete watching of a video assigned in the second instructional week before class (BI02_2) related to lower final scores. The video 2_2 was on scoring in performance assessment including analytic and holistic rubrics. Compared to other videos covering unfamiliar (e.g., 4_2, 4_3), technical (e.g., 6_2), or difficult contents (e.g., 11_2, 11_4), this video is said to be easy.

Fourth, three 'AI' variables were selected as important. Although all the three AI variables were about rather unfamiliar or difficult contents, the coefficients of the earlier videos (videos 4_2 and 4_3) were negative and that of the last week was positive (video 11_4). Specifically, AI04_3 was negative, but AC04_3 was positive. Students who completed watching video 4_3 multiple times after class obtained higher final scores, but students' unsuccessful attempts to watching it after class related to lower scores. By contrast, students' mere attempts to watch the difficult last video (AI11_4) were associated with higher scores. As was with the first video, this last video also turned out to be crucial in predicting student achievement; both AI11_4 and AC11_4 had positive coefficients.

2) STUDENT AND OTHER LEARNING BEHAVIOR VARIABLES

Irrespective of regularization methods, students' attitudes toward the course (attitudes) were the most frequently

TABLE 5. Descriptive statistics on students' video watching frequencies per video.

	I(incomplete)			C (complete)		
	Min	Mean(SD)	Max	Min	Mean(SD)	Max
B(before class)	0.00	0.19 (0.42)	1.66	0.00	0.22 (0.43)	1.15
A(after class)	0.25	1.01 (0.59)	2.19	0.25	1.07 (0.40)	1.42

Note: The mean values smaller than 1 indicate that students on average did not watch all the videos.

selected predictor, followed by completion of pre-class watching of the first video (BC01_1) and grade level (on-grade). Specifically, students' positive attitudes toward measurement and evaluation were associated with higher scores; one unit increase in the mean score related to 1.39 (Enet) and 1.67 (Mnet) increase in the final. On-grade students, who were sophomores, had 0.58 and 0.76 lower scores than off-grade students with Enet and Mnet, respectively. Student gender was also selected important. When the other variables were held constant, male students had 0.56 and 0.70 lower scores than female students with Enet and Mnet, respectively.

Among variables extracted from log data, the total number of clicks on SPSS material postings (spss.sum), the numbers of quiz-taking via mobile or PC (test.M and test.P), and the frequencies of video watching via mobile (lecture.M) were important predictors to the final scores. More clicks on SPSS postings related to higher scores. Specifically, one more click on the SPSS material was associated with 0.11 (Enet) and 0.16 (Mnet) increase in the final scores. Similarly, although students knew that the scores on quizzes did not count toward the final grade, simply taking the quizzes related to higher scores regardless of the device (mobile or PC). Interestingly, mobile test-taking contributed to the final more than PC test-taking did. The coefficients of mobile test-taking were 0.34 (Enet) and 0.50 (Mnet), while the corresponding coefficients of PC were smaller: 0.23 (Enet) and 0.30 (Mnet). Likewise, students who watched the instructional videos mobile (lecture.M) also obtained higher final scores, but lecture.P was not selected important. Specifically, one more video watched mobile was associated with score increase of 0.01 (Enet) or 0.02 (Mnet).

C. VIDEO-WATCHING PATTERNS AT AN INSTRUCTIONAL UNIT (RQ3)

The third research question was to investigate students' video watching patterns at an instructional unit level. Table 5 summarizes the descriptive statistics of BI, BC, AI, and AC. Although video watching was pre-class assignments, students on average completed watching only about 20% of the videos before class (BC mean = 0.22). Rather, they watched them after class (AC mean = 1.07). The range of students' video watching frequencies was quite wide. Some students clicked none of the videos before class (BI min = 0.00; BC min = 0.00), while others after class attempted to watch and finished watching each video as many as 2.19 (AI max) and 1.42 times (AC max), respectively.

Figure 3 shows each video's BI, BC, AI, and AC averages and 95% confidence intervals. The horizontal axis is the 34 videos in class progression (instructional weeks of 1 to 11), and the vertical axis indicates the average watching frequencies. The plot shows interesting patterns. Students completed the pre-class assignments in weeks 1 to 3, but they stopped watching assigned videos in the following weeks of 4 to 11. Particularly, BC averages of weeks 4 and 11 plummeted to nearly 0, and both AI and AC spiked during this period. While AC values were consistent, AI fluctuated.

VI. DISCUSSION

A. REGULARIZATION AND LEARNING ANALYTICS

The choice of ML techniques to employ depends on the data to analyze and research questions to answer. For example, image data used in classic face recognition [51] are noisy data, consisting of pixel information. Each pixel serves as a variable, and the relations are very complex. The sole research question is to predict face patterns, not to explain pixel variables. Considering the data characteristics and research question, models incorporating nonlinear, higher-order interactions will outperform those consisting of only main-effects. Nonlinear methods such as RF, SVM, and deep learning are well-suited for those unstructured, high-dimensional data; and will be likely to outperform linear models, particularly when enough training data are available.

Nevertheless, nonlinear methods are not a panacea to every research question [52], and they particularly have interpretation issues. Although diagnostics such as variable importance and partial dependence plots have been in use to assuage the interpretation problems, heavily predictive models are in essence complex ones, and only incomplete pictures of the results can be shown [27]. Relatedly, studies employing neural networks (e.g., [31]) or RF (e.g., [32]) did not discuss individual predictors' effects on student performance. However, educational research requires much more explanation than face recognition does. Although prediction is still important, researchers and practitioners in education will benefit from knowing what variables are related to students' academic achievement under what conditions.

Therefore, ML methods of interpretability deserve attention in learning analytics. Regularization produces explainable prediction models among ML, as it is based on linear regression and selects important predictors after shrinkage. Beyond producing interpretable models, the Enet and Mnet models of this study were comparable to RF models in terms of prediction. Likewise, multiple studies across diverse disciplines reported that linear models are comparable to RF (e.g., [53]–[55]) or even better than RF (e.g., [19], [20], [56]). These studies with ours have in common that the variables were pre-selected based on previous research. For instance, our raw data were unstructured LMS log data, but we extracted predictor candidates based on Pintrich's conceptual framework for SRL. This approach also appears to have increased the SNR (signal-to-noise ratio) of our cleaned data. Relatedly, Boulesteix *et al.* [57] and Probst *et al.* [46] recommend small and large mtry of RF for high and low

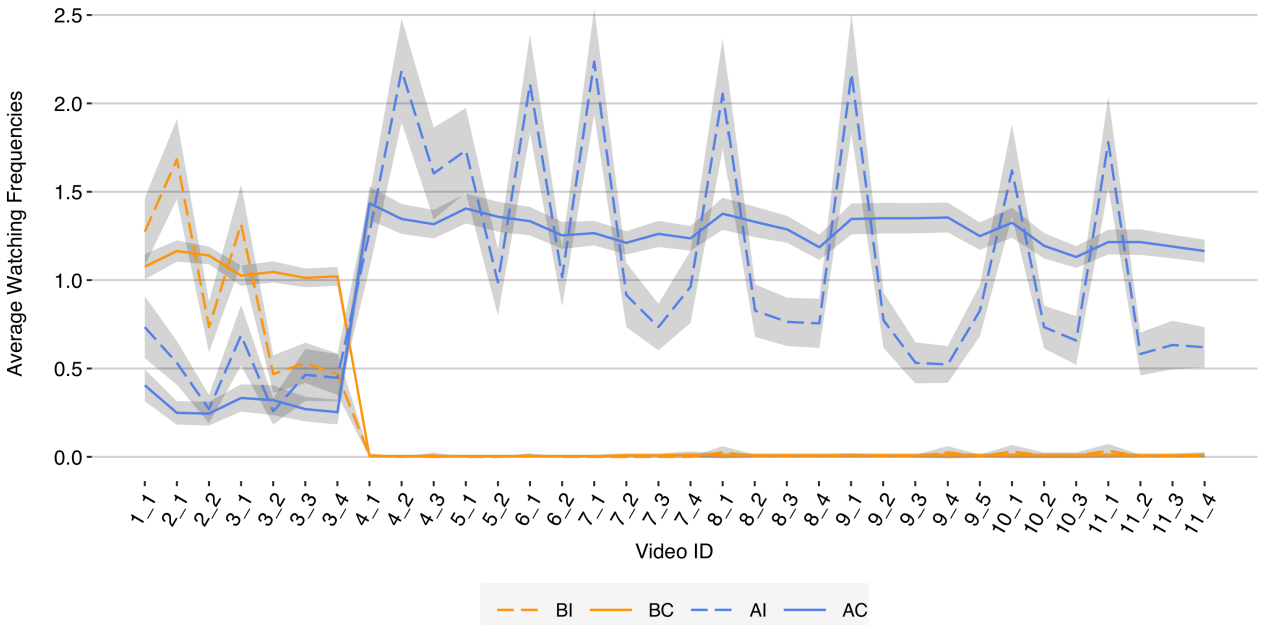

FIGURE 3. Average watching frequencies and the 95% confidence intervals of BI, BC, AI, and AC.

SNR data, respectively. Our mtry values after tuning were smaller than that of Breiman's suggestion, which can be an indication of high SNR. These altogether highlight the possible advantages of employing regularization in learning analytics.

B. IMPORTANT VARIABLES FOR PREDICTING LEARNING ACHIEVEMENT

1) VIDEO WATCHING BEHAVIORS

The prediction models could identify potential low- or high-performers as early as the first instructional week, right after the orientation week. Specifically, multiple viewings of the first video before class was a strong predictor of higher final scores, while after class viewings of the same video produced opposite results. This finding is indicative of the importance of SRL behaviors, particularly forethought and planning [30] by high performance students. Relatedly, students who attempted but failed to complete watching the second week's video multiple times before class had lower scores. Students who failed to watch the first and second week's videos before class might have had other issues not directly related to class. The significance of these variables demonstrates that procrastination and lack of persistence, a strong predictor of achievement, may be revealed at a very early stage. Thus, it appears that instructor intervention early in the semester is worth pursuing in FCs.

Watching videos completely with difficult or unfamiliar content (weeks 4, 6, and 9 to 11) multiple times after class lead to higher scores in the final exam. High performers seem to have thought that they did not perfectly understand those videos covering difficult or unfamiliar contents and thus repeated watching them after class. This result also demonstrates higher-performing students' employment of SRL behaviors. That is, higher-performing students recognized that they did not understand the contents to their content, and strategically spent extra hours on reviewing the technical or difficult videos after class.

On the contrary, unsuccessful attempts to watch videos on unfamiliar contents (e.g., Ebel and Angoff standard setting) covered in the earlier week (week 4) after class decreased the scores, while students who completed watching the same video multiple times after class increased their scores. Repeated complete viewings of challenging contents after in-class meetings indicate their self-monitoring behaviors, increased effort, and persistence. High and low performers differed in terms of how they plan and invest their efforts in relevance to the instructional context.

Findings from previous research which tend to use aggregate variables such as pre-class quiz scores [11] or self-reports on pre-class engagement [58] provide insight into factors predicting FC outcomes. Following the previous research, we included 8 aggregate variables in the prediction models, including BI (incomplete attempt before class), BC (complete watching before class), AI (incomplete attempt after class), and AC (complete watching after class). However, none of them turned out to be important. The regularization methods that this study employed (Enet and Mnet) handle multicollinear data by selecting predictors in close relations together. If the aggregate variables were important to predict student achievement, they could have been selected. This signifies the need to investigate students' behavioral data at an instructional level unit.

2) STUDENT AND OTHER LEARNING BEHAVIOR VARIABLES

Among student variables, attitudes toward the course, gender, and on-grade were important for predicting the final scores.

Particularly, students' attitudes toward the course, measured in the second week, was the most frequently selected predictor. Unlike other studies which collected students' pre-class engagement via self-reported surveys (e.g., [58]), the current study measured students' attitudes or perceptions toward the course. As expected, students who showed positive attitudes toward the course from the beginning obtained higher scores in the final. Male students obtained lower scores than female students, and on-grade students had lower scores than off-grade students. Gender and academic status have been identified as factors impacting academic outcomes [59], [60]). Particularly in college FCs, females possessed higher academic readiness than males, and older students had higher preference for FC methods than freshmen [61]. The course was intended for sophomores, but no freshmen took the classes and all the off-grade students were senior students. Test-savvy senior students might have had advantages over sophomores.

Other important variables included clicks on SPSS material postings, the number of quizzes taken via PC or mobile, and the frequencies of mobile video watching, the regression coefficients of which were all positive. These variables represent the tendency to make extra effort for mastering contents of the course, i.e., mastery goal orientation [62], whenever possible. In particular, although the quiz scores were not counted toward the final grade, the more students took the quizzes (PC or mobile), the higher their final scores were. Testing is known to be one of the most effective methods for academic success [63], and students' quiz-taking appears to relate to students' self-monitoring behaviors. Lastly, the positive relation between mobile learning and academic achievement may be due to the learner's motivation to invest more effort anytime and anywhere. This finding is in line with the systematic review by Crompton and Burke [64] regarding the effects of mobile learning on academic achievement. Although it is beyond the scope of the current study to identify the cause of this relationship, the finding of this study does suggest that providing the choice of technological platforms is necessary to address the needs of various students.

C. VIDEO WATCHING PATTERNS AT AN INSTRUCTIONAL UNIT

The results of this study demonstrate that differentiating before and after class viewing behaviors in combination with complete and incomplete viewing behaviors, provide a deeper look into the study patterns of students. Contrary to the premise of FC that students should prepare themselves for in-class meetings by watching lecture recordings in advance, many students did not complete watching lecture videos ahead. Regardless of completion, students watched the videos after class. From an instructional design perspective, these results imply that instructors might have to redesign the structure and strategies used in FCs. Randomly assigning students to summarize the lecture at the beginning of in-class meetings, allowing only students who completed pre-class assignments to open in-class materials, or setting up a reward system for completed viewings ahead of class may be strategies worth considering.

Interestingly, the pattern of after-class complete viewing was consistent and its frequency was over 1.0. Students who went through the entire video after class were likely to receive higher scores for their final exams. These results reveal the study patterns of high achieving students taking advantage of the lecture videos that FCs offer. At the same time this also implies that, FCs are similar to other forms of online classes in terms of providing lecture videos for repeated viewings, but not serving its original purpose which is enhancing in-class learning.

In contrary, incomplete after class viewings fluctuated for nearly the entire semester, with students opening the videos multiple times for the first video of each week and less than once for the rest of each week's videos. This pattern of quitting early may be due to the characteristics of these students who lack the motivation to pursue their goals or the difficulty level of this course. Whatever the cause may be, this implies the need for instructors to help these students solve the difficulties they face during the semester. Regularity of pre-class activities including video watching behaviors has been identified as important indicators for learning achievement [39]. The current study demonstrates the need to look beyond regularity and investigate how students react to changes in the instructional context as well.

Furthermore, the question of why the FC model is not working as intended may be raised. The incompatibility of students' total workload with individual FCs requiring planned effort and time, or the general lack of enthusiasm of the ordinary student after the early weeks of the semester may be the cause for very low viewings before class. Other factors may include the nature of this course, such as mandatory enrollment, and the switch from norm-referenced scoring to criterion-referenced scoring due the COVID-19 pandemic. Extremely low levels of preparation for classes imply the complexities involved in bringing about successful participation of students in FCs.

VII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Based on the findings, we propose regularization as a suitable ML method in learning analytics. Regularization particularly combined with LMS log data can explore unknown relationships among variables related to students' learning in the online environment. Beyond producing interpretable prediction models, the regularization of this study showed comparable prediction to RF. More regularization research in learning analytics will open up many possibilities for using educational big data.

Like other ML techniques, regularization may not yield the same results mainly due to data-splitting in model validation, which can be an issue when variable selection is of interest. Among 159 predictors explored, as many as 135 predictors were selected at least once and none of them was selected in all 1,000 iterations. This asserts the need to build models multiple times and employ selection counts with regularization.

However, there has been no research on selection counts in the context of learning analytics. The criteria on the selection counts deserve attention in future research. Another methodological finding is that tuning in random forest did not outperform Breiman's default, but more research is warranted on this topic to give practical guidance to researchers in learning analytics.

Findings of this empirical study are either in line with the existing literature, or present variables rarely recognized in previous studies. Variables such as gender, grade level, attitudes towards the course have been identified to have impact on academic achievement in previous studies. Variables less frequently recognized in previous studies include completion of the first instructional video before class, completion of videos of unfamiliar or difficult contents after class, or mere attempts to watch the last difficult video as well as mobile learning and non-mandatory quiz-taking. Further exploration of these variables with other potentially important predictors (e.g., students' intellectual/socioemotional development, career-related variables, family and school factors) through regularization can provide more insight into understanding students' learning from a comprehensive perspective.

This study also revealed that contrary to expectations most students did not watch videos before in-class meetings but rather completely/incompletely watched them after class. Further research on patterns of self-regulated learning behaviors in relation to pedagogical factors (e.g., motivation strategies, instructional strategies) is needed to understand how students respond to various instructional stimuli presented to them during classes.

Given the importance of out-of-class self-directed learning in FCs, we will need to make more efforts to establish stronger links between pre-class assignments and in-class team projects, which will secure students' motivation in this course or similar courses, no matter where they were in the beginning. The low level of pre-class learning from the beginning of the semester and the fluctuating pattern of incomplete video watching make us consider methods for making in-class activities and pre-class learning materials more attractive, or if FC is an effective method for them at all. Stronger links can be made not only by developing class activities heavily relying on pre-class learning, but also through technological solutions.

The methods of this study allowed us to draw a detailed picture of what actually takes place in FCs in terms of learning behaviors. Traditional studies using aggregate variables had difficulty identifying which behaviors were highly related to desirable outcomes. By utilizing ML methods in this study, we hope that methods for fine-tuning interventions, feedback, lectures, and materials to the individual can be developed with more precision. The COVID-19 pandemic accelerated the rate of digitalized learning. The use of online platforms containing various tools for distributing materials, lecturing, testing, and interacting is likely to be the norm even for late-adapters before the pandemic. As the increase in variety and amount of data collected through future developments of LMSs are expected, we anticipate to see more regularization research in learning analytics in the near future.

APPENDIX
VIDEO IDs, LABELS, AND DIFFICULTY

	Video ID	Label	Difficulty
1	01_1	formative assessment	Easy
2	02_1	performance assessment: definition	Easy
3	02_2	performance assessment: scoring	Easy
4	03_1	test construction steps	Easy
5	03_2	multiple-choice items	Easy
6	03_3	constructed-response items	Easy
7	03_4	scoring caveats	Medium
8	04_1	norm-referenced evaluation	Medium
9	04_2	criterion-referenced evaluation	Medium
10	04_3	Ebel and Angoff standard setting	Difficult
11	05_1	variables and scales	Medium
12	05_2	sampling	Medium
13	06_1	descriptive statistics	Medium
14	06_2	descriptive statistics (SPSS)	Medium
15	07_1	measuring affective domains	Easy
16	07_2	observation	Easy
17	07_3	interview	Easy
18	07_4	survey	Easy
19	08_1	item difficulty and discrimination I	Medium
20	08_2	covariance and correlation	Difficult
21	08_3	item difficulty and discrimination II	Difficult
22	08_4	item difficulty and discrimination (SPSS)	Difficult
23	09_1	introduction to reliability	Medium
24	09_2	types of reliability	Difficult
25	09_3	Cronbach's alpha	Difficult
26	09_4	standard error of measurement	Difficult
27	09_5	factors influencing reliability	Difficult
28	10_1	objectivity and reliability	Medium
29	10_2	reliability (SPSS)	Difficult
30	10_3	objectivity (SPSS)	Medium
31	11_1	content validity	Medium
32	11_2	criterion-related validity	Difficult
33	11_3	construct validity	Difficult
34	11_4	relationships between reliability and validity	Difficult

Note: Instructors rated difficulty of each video into three difficulty categories before semester started.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the Educational Data Mining Conference 2021 (Paris, France) [65]. We appreciate the anonymous reviewers of the earlier version and this paper as well as the associate editor of IEEE ACCESS.

REFERENCES

[1] F. Chen, A. M. Lui, and S. M. Martinelli, "A systematic review of the effectiveness of flipped classrooms in medical education," *Med. Educ.*, vol. 51, no. 6, pp. 585–597, Jun. 2017, doi: 10.1111/medu.13272.

[2] K. F. Hew and C. K. Lo, "Flipped classroom improves Student learning in health professions education: A meta-analysis," *BMC Med. Educ.*, vol. 18, no. 1, p. 38, Mar. 2018, doi: 10.1186/s12909-018-1144-z.

[3] J. McGivney-Burelle and F. Xue, "Flipping calculus," *PRIMUS*, vol. 23, no. 5, pp. 477–486, Apr. 2013, doi: 10.1080/10511970.2012.757571.

[4] Y. Shi, Y. Ma, J. MacLeod, and H. H. Yang, "College students' cognitive learning outcomes in flipped classroom instruction: A meta-analysis of the empirical literature," *J. Comput. Educ.*, vol. 7, no. 1, pp. 79–103, Mar. 2020, doi: 10.1007/s40692-019-00142-8.

[5] J. Bergmann and A. Sams, *Flip Your Classroom: Reach Every Student in Every Class Every Day*. Washington, DC, USA: International Society for Technology in Education, 2012, pp. 13–17.

[6] M. Bond, "Facilitating student engagement through the flipped learning approach in K-12: A systematic review," *Comput. Educ.*, vol. 151, Jul. 2020, Art. no. 103819, doi: 10.1016/j.compedu.2020.103819.

[7] A. A. ElSayed, M. Caeiro-Rodriguez, F. A. Mikic-Fonte, and M. Llamas-Nistal, "Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review," presented at the 18th World Conf. Mobile Contextual Learn, Sep. 2019. [Online]. Available: https://www.learntechlib.org/p/210600/

[8] E. Fincham, D. Gasevic, J. Jovanovic, and A. Pardo, "From study tactics to learning strategies: An analytical method for extracting interpretable representations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 59–72, Jan. 2019, doi: 10.1109/TLT.2018.2823317.

[9] M. Manso-Vazquez, M. Caeiro-Rodriguez, and M. Llamas-Nistal, "An xAPI application profile to monitor self-regulated learning strategies," *IEEE Access*, vol. 6, pp. 42467–42481, 2018, doi: 10.1109/ACCESS.2018.2860519.

[10] E. Popescu and F. Leon, "Predicting academic performance based on learner traces in a social learning environment," *IEEE Access*, vol. 6, pp. 72774–72785, 2018, doi: 10.1109/ACCESS.2018.2882297.

[11] J. Lee and H. Choi, "Rethinking the flipped learning pre-class: Its influence on the success of flipped learning and related factors," *Brit. J. Educ. Technol.*, vol. 50, no. 2, pp. 934–945, Mar. 2019, doi: 10.1111/bjet.12618.

[12] L. Macfadyen and S. P. Dawson, "Numbers are not enough: Why e-learning analytics failed to inform an institutional strategic plan," *J. Educ. Technol. Soc.*, vol. 15, no. 3, pp. 149–163, Jul. 2012. [Online]. Available: https://www.jstor.org/stable/pdf/jeductechsoci.15.3.149.pdf

[13] J. W. You, "Identifying significant indicators using LMS data to predict course achievement in online learning," *Internet Higher Educ.*, vol. 29, pp. 23–30, Apr. 2016, doi: 10.1016/j.iheduc.2015.11.003.

[14] M.-H. Cho and J. S. Yoo, "Exploring online students' self-regulated learning with self-reported surveys and log files: A data mining approach," *Interact. Learn. Environ.*, vol. 25, no. 8, pp. 970–982, Nov. 2017, doi: 10.1080/10494820.2016.1232278.

[15] V. C. Smith, A. Lange, and D. R. Huston, "Predictive modeling to forecast Student outcomes and drive effective interventions in online community college courses," *Online Learn.*, vol. 16, no. 3, pp. 51–61, Jun. 2012.

[16] A. Sheshadri, N. Gitinabard, C. F. Lynch, T. Barnes, and S. Heckman, "Predicting student performance based on online study habits: A study of blended courses," presented at the Int. Conf. Educ. Data Mining, Apr. 2019.

[17] J. E. Yoo, "TIMSS 2011 student and teacher predictors for mathematics achievement explored and identified via elastic net," *Frontiers Psychol.*, vol. 9, p. 317, Mar. 2018, doi: 10.3389/fpsyg.2018.00317.

[18] J. Beemer, K. Spoon, L. He, J. Fan, and R. A. Levine, "Ensemble learning for estimating individualized treatment effects in student success studies," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 3, pp. 315–335, Sep. 2018, doi: 10.1007/s40593-017-0148-x.

[19] J. E. Yoo and M. Rho, "Exploration of predictors for Korean teacher job satisfaction via a machine learning technique, group mnet," *Frontiers Psychol.*, vol. 11, p. 441, Mar. 2020, doi: 10.3389/fpsyg.2020.00441.

[20] J. E. Yoo and M. Rho, "Large-scale survey data analysis with penalized regression: A Monte Carlo simulation on missing categorical predictors," *Multivariate Behav. Res.*, pp. 1–29, Mar. 2021, doi: 10.1080/00273171.2021.1891856.

[21] J. Liu, G. Liang, K. D. Siegmund, and J. P. Lewinger, "Data integration by multi-tuning parameter elastic net regression," *BMC Bioinf.*, vol. 19, no. 1, p. 369, Oct. 2018, doi: 10.1186/s12859-018-2401-1.

[22] C. Zeng, D. C. Thomas, and J. P. Lewinger, "Incorporating prior knowledge into regularized regression," *Bioinformatics*, vol. 37, no. 4, pp. 514–521, May 2021, doi: 10.1093/bioinformatics/btaa776.

[23] M. A. Nabian and H. Meidani, "Physics-driven regularization of deep neural networks for enhanced engineering design and analysis," *J. Comput. Inf. Sci. Eng.*, vol. 20, no. 1, Feb. 2020, Art. no. 011006, doi: 10.1115/1.4044507.

[24] Y. Zhang, R. E. Minchin, Jr., and D. Agdas, "Forecasting completed cost of highway construction projects using LASSO regularized regression," *J. Construct. Eng. Manage.*, vol. 143, no. 10, Oct. 2017, Art. no. 4017071, doi: 10.1061/(ASCE)CO.1943-7862.0001378.

[25] R. Bertolini, "Evaluating performance variability of data pipelines for binary classification with applications to predictive learning analytics," Ph.D. dissertation, Dept. Appl. Math. Eng., Stony Brook Univ., New York, NY, USA, 2021.

[26] J.-Y. Wu, "Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement," *Comput. Educ.*, vol. 163, Apr. 2021, Art. no. 104066, doi: 10.1016/j.compedu.2020.104066.

[27] N. Bosch, "Identifying supportive student factors for mindset interventions: A two-model machine learning approach," *Comput. Educ.*, vol. 167, Jul. 2021, Art. no. 104190, doi: 10.1016/j.compedu.2021.104190.

[28] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educ. Psycholog.*, vol. 25, no. 1, pp. 3–17, Jan. 1990, doi: 10.1207/s15326985ep2501_2.

[29] J. Broadbent and W. L. Poon, "Self-regulated learning strategies & Academic achievement in online higher education learning environments: A systematic review," *Internet Higher Educ.*, vol. 27, pp. 1–13, Oct. 2015, doi: 10.1016/j.iheduc.2015.04.007.

[30] P. R. Pintrich, "A conceptual framework for assessing motivation and self-regulated learning in college students," *Educ. Psychol. Rev.*, vol. 16, no. 4, pp. 385–407, Dec. 2004, doi: 10.1007/s10648-004-0006-x.

[31] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Analyzing learners behavior in MOOCs: An examination of performance and motivation using a data-driven approach," *IEEE Access*, vol. 6, pp. 73669–73685, 2018, doi: 10.1109/ACCESS.2018.2876755.

[32] P. M. Moreno-Marcos, T.-C. Pong, P. J. Munoz-Merino, and C. D. Kloos, "Analysis of the factors influencing learners' performance prediction with learning analytics," *IEEE Access*, vol. 8, pp. 5264–5282, 2020, doi: 10.1109/ACCESS.2019.2963503.

[33] E. Fincham, D. Gasevic, J. Jovanovic, and A. Pardo, "From study tactics to learning strategies: An analytical method for extracting interpretable representations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 59–72, Jan. 2019, doi: 10.1109/TLT.2018.2823317.

[34] R. Cerezo, M. Esteban, M. Sánchez-Santillán, and J. C. Núñez, "Procrastinating behavior in computer-based learning environments to predict performance: A case study in moodle," *Frontiers Psychol.*, vol. 8, Aug. 2017, Art no. 1403, doi: 10.3389/fpsyg.2017.01403.

[35] J. Jovanovic, N. Mirriahi, D. Gašević, S. Dawson, and A. Pardo, "Predictive power of regularity of pre-class activities in a flipped classroom," *Comput. Educ.*, vol. 134, pp. 156–168, Jun. 2019, doi: 10.1016/j.compedu.2019.02.011.

[36] T. Hastie, R. Tibshirani, and J. Freedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, pp. 51–52.

[37] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[38] J. Huang, P. Breheny, S. Lee, S. Ma, and C.-H. Zhang, "The Mnet method for variable selection," *Statistica Sinica*, vol. 26, no. 3, pp. 903–923, Jul. 2016. [Online]. Available: https://www.jstor.org/stable/24721259

[39] T. Hastie, J. Qian, and K. Tay. (2021). *An Introduction to Glmnet*. Accessed: Nov. 10, 2021. [Online]. Available: https://cloud.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf

[40] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, Nov. 2003, doi: 10.1093/bioinformatics/btg308.

[41] U. Meinshausen and P. Bühlmann, "Stability selection," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 72, no. 4, pp. 417–473, 2010, doi: 10.1111/j.1467-9868.2010.00740.x.

[42] T. Hastie, R. Tibshirani, and J. Freedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, pp. 241–242.

[43] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, "Exact post-selection inference, with application to the lasso," *Ann. Statist.*, vol. 44, no. 3, pp. 907–927, 2016, doi: 10.1214/15-AOS1371.

[44] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Statist. Comput.*, vol. 25, no. 2, pp. 173–187, 2015, doi: 10.1007/s11222-013-9424-2.

[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[46] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 3, Jan. 2019, Art. no. e1301, doi: 10.1002/widm.1301.

[47] B. A. Goldstein, E. C. Polley, and F. B. Briggs, "Random forests for genetic association studies," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, Jul. 2011, doi: 10.2202/1544-6115.1691.

[48] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

[49] P. Probst and A. L. Boulesteix, "To tune or not to tune the number of trees in random forest," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1–18, Apr. 2018. [Online]. Available: https://www.jmlr.org/papers/volume18/17-269/17-269.pdf

[50] J. Huang, S. Ma, and C. H. Zhang, "Adaptive Lasso for sparse high-dimensional regression models," *Stat. Sinica*, vol. 18, no. 4, pp. 1603–1618, 2008. [Online]. Available: https://www.jstor.org/stable/24308572

[51] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996, doi: 10.1109/34.531802.

[52] T. Gervet, K. Koedinger, J. Schneider, and T. Mitchell, "When is deep learning the best approach to knowledge tracing," *J. Educ. Data Mining*, vol. 12, no. 3, pp. 31–54, Dec. 2020, doi: 10.5281/zenodo.4143614.

[53] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[54] M. Liu, S. Hu, Y. Ge, G. B. M. Heuvelink, Z. Ren, and X. Huang, "Using multiple linear regression and random forests to identify spatial poverty determinants in rural China," *Spatial Statist.*, vol. 42, Apr. 2021, Art. no. 100461, doi: 10.1016/j.spasta.2020.100461.

[55] M. R. Pahlavan-Rad, K. Dahmardeh, M. Hadizadeh, G. Keykha, N. Mohammadnia, M. Gangali, M. Keikha, N. Davatgar, and C. Brungard, "Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran," *CATENA*, vol. 194, Nov. 2020, Art. no. 104715, doi: 10.1016/j.catena.2020.104715.

[56] P. F. Smith, S. Ganesh, and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience," *J. Neurosci. Methods*, vol. 220, no. 1, pp. 85–91, Oct. 2013, doi: 10.1016/j.jneumeth.2013.08.024.

[57] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 2, no. 6, pp. 493–507, 2012, doi: 10.1002/widm.1072.

[58] J. Lee, T. Park, and R. O. Davis, "What affects learner engagement in flipped learning and what predicts its outcomes?" *Brit. J. Educ. Technol.*, pp. 1–18, Nov. 2018, doi: 10.1111/bjet.12717.

[59] U. Kessels, A. Heyder, M. Latsch, and B. Hannover, "How gender differences in academic engagement relate to students' gender identity," *Educ. Res.*, vol. 56, no. 2, pp. 220–229, Apr. 2014, doi: 10.1080/00131881.2014.898916.

[60] D. K. Wentworth and J. H. Middleton, "Technology use and academic performance," *Comput. Educ.*, vol. 78, pp. 306–311, Sep. 2014, doi: 10.1016/j.compedu.2014.06.012.

[61] Y. Hao, "Exploring undergraduates' perspectives and flipped learning readiness in their flipped classrooms," *Comput. Hum. Behav.*, vol. 59, pp. 82–92, Jun. 2016, doi: 10.1016/j.chb.2016.01.032.

[62] C. Ames and J. Archer, "Achievement goals in the classroom: Students' learning strategies and motivation processes," *J. Educ. Psychol.*, vol. 80, no. 3, pp. 260–267, Sep. 1988.

[63] K. A. Rawson and J. Dunlosky, "When is practice testing most effective for improving the durability and efficiency of student learning?" *Educ. Psychol. Rev.*, vol. 24, no. 3, pp. 419–435, Sep. 2012, doi: 10.1007/s10648-012-9203-1.

[64] H. Crompton and D. Burke, "The use of mobile learning in higher education: A systematic review," *Comput. Educ.*, vol. 123, pp. 53–64, Aug. 2018, doi: 10.1016/j.compedu.2018.04.007.

[65] J. E. Yoo and M. Rho, "LMS log data analysis from fully-online flipped classrooms: An exploratory case study via regularization," presented at the Int. Conf. Educ. Data Mining, Jul. 2021. [Online]. Available: https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_80.pdf

JIN EUN YOO received the B.A. degree from Seoul National University and two M.S. degrees in educational psychology and applied statistics and the Ph.D. degree in measurement, statistics, and research methodology from Purdue University, West Lafayette, IN, USA. She was a Psychometrician with Pearson, Austin, TX, USA, and a Research Scholar with the Department of Computer Science, San Francisco State University. She has been a Professor with the Korea National University of Education. Her research interests include educational data mining, statistical learning, and missing data analysis. She is a member of the American Educational Research Association and the National Council on Measurement in Education. She currently serves as an Associate Editor for *Frontiers in Psychology: Quantitative Psychology and Measurement* and *Innovation and Education*.

MINJEONG RHO received the bachelor's degree from the Gongju National University of Education and master's and Ph.D. degrees in curriculum from the Korea National University of Education. She has been a Lecturer with the Korea National University of Education and an Elementary School Teacher. Her research interests include educational evaluation, curriculum, educational data mining, and machine learning.

YEKYUNG LEE received the bachelor's and master's degrees from Seoul National University (SNU) and the Ph.D. degree in educational technology from Purdue University. She worked as a Senior Researcher with the Center for Human Resource Development, SNU. She joined the Sogang Graduate School of Education, Sogang University, as a Faculty Member. Her research interests include instructional methods based on social psychology, instructional design for developing thinking skills, and integrating technology for student centered learning. She was a member of the Presidential Committee of the 4th Industrial Revolution in Korea and recommended plans and policies for the future of education in Korea. She is also a Committee Member of the Korean Society for Educational Technology and the Korean Association for Educational Information and Media.