Received: 31 August 2021 | Revised: 22 October 2021 | Accepted: 17 November 2021

DOI: 10.1002/cae.22479

RESEARCH ARTICLE

WILEY

Development of a student engagement score for online undergraduate engineering courses using learning management system interaction data

Javeed Kittur | Jennifer Bekki | Samantha Brunhaver

Engineering Education Systems and Design, Arizona State University, Mesa, Arizona, USA

Correspondence
Javeed Kittur, Engineering Education Systems and Design, Arizona State University, Mesa, AZ 85212, USA.
Email: jkittur@asu.edu and javedkittur89@gmail.com

Funding information
National Science Foundation, Grant/Award Number: 1825732

Abstract

Although researchers agree that student engagement in online courses is a function of time dedicated to course-related activities, there is little consensus about the best way to quantify the construct. This study introduces a measure for undergraduate engineering students' engagement in online courses using their interactions with their online course learning management system (LMS). Data from 81 courses offered by three fully online, undergraduate engineering degree programs generated a total of 3848 unique student–course combinations (approximately 2.7 million rows of LMS interaction data), to which we applied a five-step process to calculate a single score representing student LMS engagement. First, we converted the students' LMS interaction data into a set of *natural* features representing the time they spent per 3-day period on various course elements, such as quizzes, assignments, discussion forums, and so forth, and how these times changed across the duration of the course. We then used the natural features to derive 216 *relative* features describing deviations from typical interaction patterns among students in the same course. Next, we conducted association rule mining on a training portion of the data set to generate rules separately describing the behavior of students who completed the course (completers) and those who chose to drop early (leavers). The rules generated were applied to students from the testing portion of the data set to compute the percentage of unique rules met by completers and leavers. Finally, the mathematical difference between the percentages of completer and leaver rules met by each student was found to be the best measure of student engagement.

KEYWORDS
association rule mining, completers, engagement score, leavers, online engineering

1 | INTRODUCTION

Online education is rapidly expanding due to its accessibility, scalability, and flexibility [5,44]. One of the major challenges in online courses is student course-level attrition, which is higher in the online format than in face-to-face courses [11,22,45]. Researchers have tried to address higher attrition in online courses by investigating its probable causes. For example, Hart [24] identified motivation, online learning satisfaction, sense of belonging in the community, peer and family support, communication with the instructor, and time management skills as factors influencing students' decision to persist in online courses. Other important factors in students' successful completion of online courses have included students' prior academic achievement, previous information technology training, continuous academic enrollment, and financial assistance [43].

Researchers have also predicted online students' course persistence using data describing the students' patterns of interaction with their online course [14,17,25,35,42,54]. For example, Shelton et al. [46] identified students at risk of dropping their online course using student–teacher and student–student interaction data, where the frequency of online interactions proved to better indicate student persistence and success than did the length of interactions. Aguiar et al. [4] predicted persistence using first-year engineering students' electronic portfolios, extracting information about their course engagement through their reflections about engineering advising, project updates, and engineering exploration throughout the course. Using attributes related to student activities such as assignment skips, assessment performance, and video skips and lags to predict student dropout in online courses, Halawa et al. [23] were able to successfully flag 40%–50% of students who dropped out of the course while they were still enrolled. Finally, a study by Morris and Finnegan [36] student attribute data and student course interaction data to predict students' course-level persistence decisions in separate studies.

Each of the studies above underscores the potential to use data related to students' activities in online courses to predict students' persistence decisions. This paper similarly presents evidence supporting the development and efficacy of a student engagement measure based on the student–learning management system (LMS) data interaction patterns that uniquely identify course leavers and completers in online undergraduate engineering courses. We focus on online undergraduate engineering students specifically, given the steadily increasing number of online courses and programs for undergraduate engineering students over the last decade [31,44,57] and the potential for greater student attrition due to the difficulties of replicating in the online formal typical aspects of the undergraduate engineering experience [8,21]. This study is part of a larger National Science Foundation-funded study to develop and evaluate a theoretical model for online undergraduate engineering student persistence by combining student attribute and LMS interaction data [13]. A summary of the literature on student engagement in online courses is provided next.

2 | STUDENT ENGAGEMENT IN ONLINE COURSES

Student engagement is a construct widely considered in educational research, in both face-to-face and online modalities, due to its demonstrated correlation with several positive student outcomes, including course level persistence [38,53]. While some studies have focused on cognitive measures of student engagement such as students' motivations and strategies for learning [41], others have operationalized engagement as student effort toward educationally advancing activities [10,12,16,20] and interaction with classmates, instructors, and the courses themselves [19]. A growing body of work within this category uses learning analytics to track student engagement indicators such as the number of assignments completed, discussion board messages posted, quizzes taken, and emails written [9,28,37,48,52]. For example, Bote-Lorenzo and Gómez-Sánchez [10] calculated students' engagement scores by averaging the percentages of assignments submitted, exercises completed, and lecture videos watched through the students' course LMS and the change in students' engagement scores as the difference in percentages completed between consecutive units in the course. Yet more studies have correlated learning analytics-based measures of student engagement with student persistence. In one study, Balakrishnan and Coetzee [7] used students' interactions with their Massive Open Online Courses (MOOCs) to predict their retention in the MOOC. In another study, Kizilcec et al. [30] used students' patterns of interactions with their course LMS to predict students' engagement type (i.e., completing, auditing, sampling, or disengaging from the course) which they proposed educators could use as a warning system to identify students at risk of dropping the course.

The amount of time spent on LMS activities can help understand student engagement in online courses and time can be studied using either natural or relative reference frames. The natural reference frame refers to an individual's time spent on LMS-related activities and the change in individual's time spent on LMS-related activities over a certain period and the relative reference frame refers to the individual's time spent on LMS-related activities as compared with their classmates [53,55]. Few studies [26,47,56] consider how student engagement varies over time and relative to one's peers, despite evidence that student engagement is a function of course norms [16]. Researchers lack a measure of online student engagement they can confidently utilize in their work that captures the relative reference frames.

This paper provides full details supporting our methodology to create a numerical value describing the construct of student engagement in online undergraduate engineering education. We begin addressing this goal by exploring how LMS interaction data can be used to compute student engagement scores within online undergraduate engineering courses. The following sections fully document our data set and methodology used to create a numerical value describing this construct. Our analysis offers researchers in the educational data mining space a novel approach to conduct their own investigations related to online student engagement, an important construct to studying student persistence in online courses.

3 | DATA SET

The data set for this study comes from 81 courses offered by three fully online, ABET-accredited undergraduate engineering degree programs at a large, public, southwestern university between Fall 2018 and Spring 2020. Nine courses were from electrical engineering, 35 were from engineering management, and 37 were from software engineering. All courses were 7.5 weeks in duration and used Canvas as the LMS platform. We collected approximately 2.7 million rows of LMS interaction data from 3848 unique student–course combinations. Unique student–course combinations were considered as students could be enrolled in more than one course. About 90% of student–course combinations came from students who persisted in the course to its completion. Table 1 summarizes the data set in terms of the number of courses from each program and the number of persisting and nonpersisting students for each 7.5-week period of data collection. Table 2 summarizes the student enrollment data across three degree programs based on the different course levels: introductory (100 level courses), intermediate (200 level courses), advanced intermediate (300 level courses), and advanced (400 level courses). Approximately, 17% of the total courses belong to the introductory, 30% to the intermediate, 34% to the advanced intermediate, and 19% to advanced level courses.

Each row of LMS data represents a different student interaction with their course LMS, whether navigating to a particular type of page by clicking on a link (such as to quizzes, assignments, discussion forums, modules, wiki pages, attachments, grades, the syllabus, or announcements) or submitting quizzes and assignments. Table 3 describes each activity type considered in this study. Table 4 illustrates the raw structure of the data set with deidentified student IDs and course IDs, this table has been reproduced from the previously published work [29]. The raw data includes the following elements: student ID (student_id), course ID (course_id), time of the event (eventtime), type of the event (eventtype), action related to an event (Action), activity type (object_name) and student enrolment status in the course (enrl_status).

4 | PROCEDURE AND RESULTS

4.1 | Feature creation

The graphical representation of the process used in preparing the data by creating and selecting features required to conduct association rule mining (ARM) analysis is described in Figure 1 and explained in detail in this section. We used the students' LMS interaction data to create 2161 *natural* features for each unique student–course combination. The natural features represent one of two categories of activity. First, they represent a student's time spent on LMS-related activities and include time spent on quizzes, assignments, discussion forums, wiki pages, attachments, modules, the syllabus, grades, announcements, and the LMS overall. Second, natural features also represent the raw number of quiz and assignment submissions by a student. Each natural feature was calculated over consecutive 3-day windows; for example, "time spent on quizzes" was

TABLE 1 Student enrollment data across different sessions

#	Session	Number of courses: Electrical engineering	Engineering management	Software engineering	Persisting students	Nonpersisting students
1	Fall-B 2018	1	3	0	156	17
2	Spring-A 2019	1	5	8	611	56
3	Spring-B 2019	1	6	7	581	82
4	Fall-A 2019	2	8	7	727	83
5	Fall-B 2019	3	5	6	675	79
6	Spring 2020	1	8	9	717	64

TABLE 2 Student enrollment data based on course levels across degree programs

Degree program	Course level: Introductory	Intermediate	Advanced intermediate	Advanced
Electrical engineering	663	288	-	-
Engineering management	-	-	509	407
Software engineering	-	883	775	323

TABLE 3 Description of the activity types

#	Activity Type	Description
1	Quizzes	Student submitting a quiz, student navigating to a quiz
2	Assignments	Student submitting an assignment, student modifying an assignment, student navigating to an assignment
3	Discussion forum	Student posting a message, student navigating to a discussion thread
4	Wiki pages	Student navigating to a wiki page
5	Attachments	Student navigating to an attachment
6	Modules	Student navigating to modules
7	Syllabus	Student navigating to syllabus
8	Grades	Student navigating to grades
9	Announcements	Student navigating to announcements

calculated across each 3-day period in the course (i.e., Days 1–3, Days 4–6, etc.) The length of 3 days, also referred to as the "analysis window length," or just "window length" was selected because it allowed us to detect the students' LMS temporal patterns as students may choose different times and days to work on the different tasks in the course. The 3-days data will be sufficient to analyze students' temporal patterns as considering more than 3 days as an analysis window period in a 7.5-week course could gloss over important details. The first analysis window for each course was eliminated because it corresponded with the university's semesterly course drop deadline (i.e., students can drop the class during the first 3 days without penalty). After removing this first analysis window of data, 16 analysis windows of data for each course remained. Table 5 shows how the sample data were structured [29]. The columns represent the student's time spent on quiz (t_{quiz}), assignment ($t_{\text{assignment}}$), discussion forum (t_{dforum}), wiki pages (t_{wiki}), attachments (t_{attach}), modules (t_{modules}), course syllabus (t_{syllabus}), course grades (t_{grades}), and student's course status in a given analysis window.

The broader aim of this study was to develop a numerical representation of student engagement, which is known to be a function of course norms [16]. Correspondingly, from the natural features, *relative* features, which compare LMS interaction activities of each student to the "norms" for others in their same course, were calculated. Table 6 lists all the relative features utilized in the study and includes, for example, a feature describing the difference between an individual student's time spent and the average time spent for all students in the class during the analysis windows. In total, 216 relative features describing change over time and deviations from typical LMS-interaction patterns among students in the same course, were generated. Of note is that these features, shown in Table 6, are not temporal features capturing the change in an individual student's behavior over time, but features that describe the difference between an individual student's activities and those of the "norms" within the class.

Calculating the relative features required specifying the number of analysis windows over which each relative feature would be calculated and selecting which particular analysis windows during the duration of data collection would serve as the basis of their calculation. This is an important step to meet our analysis window as we do not wish to include students who have not spent enough time and dropped from the course. Given the fact that the total percentage of course leavers was so small in comparison with the course completers, we were careful in selecting the length of the analysis window such that it captures the students' relevant behavior and to not lose a greater number of dropping students from our data set. We arranged the percentage of dropped students considering multiple analysis window lengths and we decided to use three analysis windows data. We chose to calculate relative features based on three consecutive analysis windows (e.g., analysis windows 1–3, analysis windows 2–4, analysis windows 3–5, etc.) because it was the minimum number necessary to calculate our variance-related relative features (see Table 6) while still yielding the maximum number of students who dropped in our data set during each analysis period, which was helpful in discriminating between the behavior of course leavers from that of course completers.

To discriminate the behavior of course completers and leavers, it is important to determine which analysis windows to be considered such that the relevant data

TABLE 4 Structure of the raw data [29]

Eventtime	Student_id	Course_id	Eventtype	Action	object_name	enrl_status
10/10/2018 9:21:33	A	2018FallB	NavigationEvent	NavigatedTo	quizzes:quiz	ENRL
10/15/2018 9:22:18	A	2018FallB	NavigationEvent	NavigatedTo	Attachment	ENRL
10/11/2018 19:54:17	B	2018FallB	NavigationEvent	NavigatedTo	Syllabus	ENRL
10/16/2018 15:55:03	B	2018FallB	AssessmentEvent	Submitted	-	ENRL
10/22/2018 10:06:53	C	2018FallB	NavigationEvent	NavigatedTo	Modules	ENRL
10/22/2018 17:11:47	C	2018FallB	NavigationEvent	NavigatedTo	Grades	ENRL
10/13/2018 23:05:59	D	2018FallB	AssignableEvent	Submitted	-	WDRW
10/16/2018 23:45:24	E	2018FallB	Event	Modified	-	WDRW
10/24/2018 0:00:55	F	2018FallB	NavigationEvent	NavigatedTo	announcements	WDRW

Abbreviations: ENRL, student remained enrolled in the course; WDRW, student withdrew from the course.

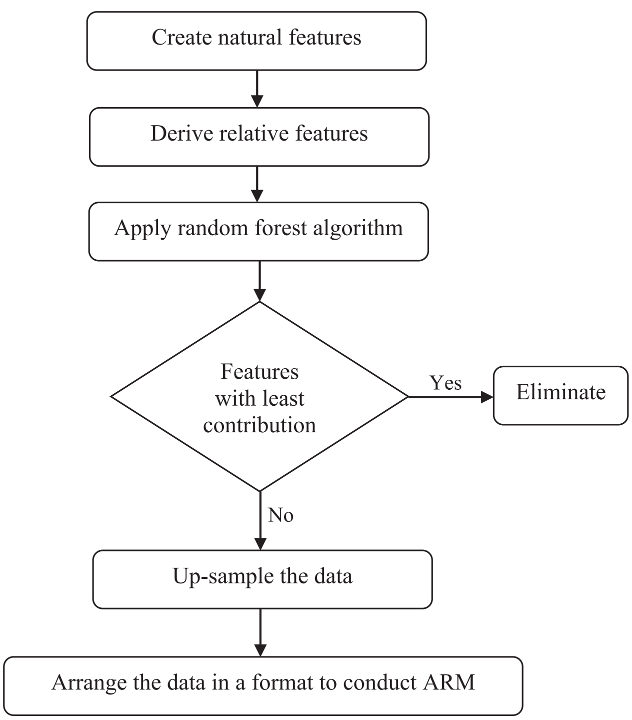

FIGURE 1 Preparation of data required to conduct association rule mining (ARM)

is available for the analysis. We also assumed that while the features for persisting students would be non-distinctive for any analysis period during the course, the period just before a student drops would include the most distinctive feature across the duration of the course for leavers. We, thus, used the last three analysis windows before a student's withdrawal from the course as the analysis period for leavers and randomly selected three consecutive analysis windows for course completers to create the relative features [29].

4.2 | Feature selection

Once the relative features were developed, we used the feature selection part of the random forest algorithm [50] to identify features that uniquely distinguish course completers from course leavers. We randomly divided into two data sets of 31 courses (data set 1) and 32 courses (data set 2), to verify the stability of selected features. Each set of features was arranged in descending order according to their random forest Gini index, the higher of which signifies the greater importance of a feature in distinguishing course completers from course

TABLE 5 Structure of the data with sample natural features in a particular analysis period [29]

Student	t_{quiz}	$t_{assignment}$	t_{dforum}	t_{wiki}	t_{attach}	$t_{modules}$	$t_{syllabus}$	t_{grades}	Status
A	57.36	0.422	0.383	278.5	193.1	111.9	4.31	3.80	ENRL
B	15.01	0.266	0.000	30.00	54.43	0.000	0.46	0.00	ENRL
C	18.81	0.100	2.450	239.7	291.1	138.2	0.01	0.18	ENRL
D	9.960	0.160	1.580	0.000	91.13	0.760	0.01	0.55	ENRL
E	48.68	0.850	1.010	184.8	32.03	1.410	0.00	0.52	ENRL
F	93.00	0.000	0.230	5.580	27.88	90.08	2.36	0.00	ENRL
G	9.580	4.130	0.570	92.50	88.91	61.75	3.35	0.28	WDRW
H	2.730	0.100	0.060	1.460	6.500	0.230	2.30	0.00	WDRW
I	109.8	0.420	0.570	227.8	16.95	183.1	0.00	0.52	WDRW
J	0.000	0.000	2.130	0.030	94.60	1.210	0.01	0.00	WDRW

Abbreviations: ENRL, student remained enrolled in the course; WDRW, student withdrew from the course.

leavers relative to other features. Table 7 shows the top 30 features selected using the feature selection process from each data set, grouped based on their associated LMS interaction activity type (e.g., quiz submission, time spent looking at grades, etc.) For example, features related to quiz submissions appeared six times in the top 30 features selected from data set 1 and five times in the top 30 features selected from data set 2. The purpose of selecting top features was to understand which relative features related to the different LMS-activity types are relatively more important in distinguishing course completers from course leavers. Relative features related to the syllabus, discussion forums, and announcements did not appear in the top 30 features selected for either data set and were removed from further analysis, reducing the number of relative features to 162. Readers are directed to Reference [29] for more details on the creation of the natural and relative features.

4.3 | Association rule mining

The process used in conducting ARM analysis is graphically presented in Figure 2 and more details about this process are described in this section. With the final 162 relative features, ARM was used to generate rules uniquely describing completers and leavers. ARM discovers hidden relationships among variables in large data sets using association rules a→b, where "a" is the antecedent of the rule, "b" is the consequent [2,32,482,33,49]. The rule a→b indicates the likelihood that a specific student's activity containing relative features in "a" will tend to include the student's persistence decision (yes/no) in "b." In this study, N refers to the set of total students with unique identifiers $\{ID_1, ID_2, ID_3, ..., ID_N\}$, "a" refers to the set of Z relative features $\{F_1, F_2, F_3, ..., F_Z\}$, and "b" refers to students' decision to persist ("1") or not persist ("0") in their online course. Table 8 illustrates the format required for data to run in ARM, where rows represent transactions (students) and columns represent the itemset "a + b" (relative features and persistence) [1,3]. For example, the first row identifies a student with student ID-1 who persisted in the course (Persistence = 1), with a low engagement rating (1) on relative features F_1 and F_z, high engagement rating (3) on relative features F_2 and F_3.

ARM requires the discretization of continuous data, which the relative features describing student engagement in our data set were. Approaches to discretize data for use in ARM include dichotomizing values based on whether it is above or below a certain threshold, dividing data into equal-sized bins, and using quartiles to assign data to different categories [6,7,34,51]. We divided the data for each relative feature, F_i, into three bins before initiating ARM. The first bin had data points less than or equal to the first quartile (Q_1), which were assigned a value of "low engagement" (LOW) relative to the average student in the course. The second bin had data points greater than the first quartile (Q_1) and less than or equal to the third quartile (Q_3), which were assigned a value of "medium engagement" (MED) relative to the average student in the course. The last and third bin had data points greater than the third quartile (Q_3), which were assigned a value of "high engagement" (HIGH) relative to the average student in the class. In Table 8, 1 = LOW, 2 = MED, and 3 = HIGH. For the third (F3) and fifth (F5) relative features (features related to the difference between an individual student and the student with maximum time/number of submissions in the class), the interpretation is slightly different from the other relative features. For features F3 and F5, a value

TABLE 6 Relative features notation and representation [29]

Feature #	Feature description and mathematical representation
	Notations: n_{jk}—Number of students in course j in analysis period k M_{ijk}—Number of submissions by student i in course j in analysis period k G_{ijk}—Time spent or number of submissions by student i in course j in analysis period k. D_{ij}—Duration of the course considered for a student i and course j N—number of windows
F1	Difference between an individual student's time spent and the average time spent for all students in the class, in a particular analysis period $G_{ijk} - \frac{\sum_{i \in n_{jk}} G_{ijk}}{n_{jk}} \ \forall \quad k \in D_{ij}$
F2	Difference between an individual student's change in time spent and the average change in time spent for all students in the class, in a particular analysis period $(G_{ijk} - G_{ijk'}) - \left[\frac{\sum_{i \in n_{jk}} (G_{ijk} - G_{ijk'})}{n_{jk}}\right]$ $\forall \quad k, k' \in D_{ij} \quad \text{and} \quad k < k'$
F3	Difference between the maximum change in time spent for all students in the class and an individual student's change in time spent, in a particular analysis period $\max_{i \in j} (G_{ijk} - G_{ijk'}) - (G_{ijk} - G_{ijk'})$ $\forall \quad k, k' \in D_{ij} \quad \text{and} \quad k < k'$
F4	Difference between an individual student's change in time spent and the minimum change in time spent for all students in the class, in a particular analysis period $(G_{ijk} - G_{ijk'}) - \min_{i \in j} (G_{ijk} - G_{ijk'})$ $\forall \quad k, k' \in D_{ij} \quad \text{and} \quad k < k'$
F5	Difference between the maximum time spent by a student in the class and the time spent by an individual student, in a particular analysis period $\max_{i \in j,k} (G_{ijk}) - G_{ijk} \quad \forall \quad i \in j, k \quad \text{and} \quad k \in D_{ij}$
F6	Difference between the time spent by an individual student and the minimum time spent by a student in the class, in a particular analysis period $G_{ijk} - \min_{i \in j} (G_{ijk}) \ \forall \quad i \in j, k \quad \text{and} \quad k \in D_{ij}$
F7	Difference between the variance of an individual student's time spent and the average variance of time spent for all students in the class across three different windows $\frac{1}{N-1} \sum_{k=1}^{N} \left[G_{ijk} - \frac{\sum_{i=1}^{n_{jk}} G_{ijk}}{n_{jk}}\right]^2 - \frac{\frac{1}{N-1} \sum_{k=1}^{N} \left[G_{ijk} - \frac{\sum_{i=1}^{n_{jk}} G_{ijk}}{n_{jk}}\right]^2}{n_{jk}}$ $\forall \quad i \in j, k \quad \text{and} \quad N \in (3 \ \text{to} \ 15) \ \text{and} \quad k \in D_{ij}$
F8	Difference between the variance of time spent by an individual student and the minimum variance of the time spent by a student in the class across different windows. $\frac{1}{N-1} \sum_{k=1}^{N} \left[G_{ijk} - \frac{\sum_{i=1}^{n_{jk}} G_{ijk}}{n_{jk}}\right]^2$ $- \min \left\{\frac{1}{N-1} \sum_{k=1}^{N} \left[G_{ijk} - \frac{\sum_{i=1}^{n_{jk}} G_{ijk}}{n_{jk}}\right]^2\right\}$ $\forall \quad i \in j, k \quad \text{and} \quad N \in (3 \ \text{to} \ 15) \ \text{and} \quad k \in D_{ij}$

LOW represents that a student's engagement was relatively more than that of a student with a value HIGH for feature types. This is because if the difference between a student's score and the maximum score in the class is smaller, it implies that the student's score was nearer to the maximum score in the class than if the difference was greater.

Once generated, the association rules for this study were mined using the *apriori* algorithm of the *arules* package in the statistical software R [39]. First, we split the discretized data into a training data set (80%) and a testing data set (20%) and conducted ARM on the training data set to generate rules capturing the behavior of course completers and course leavers, separately. A syntactic constraint restricts the items that appear in a rule [50], such as understanding how restricting items in the consequent affects the set of items in the antecedent or vice versa. Syntactic constraints were placed on the consequent of each rule, as we were interested in identifying unique rules for students who persisted and students who dropped the course, respectively. We generated the rules for course completers by fixing the syntactic constraint on the consequent to "1," which looked like {set of relative features} → {persistence = HIGH}, and the rules for course leavers by fixing the syntactic constraint on the consequent to "0," which appeared as {set of relative features} → {persistence = LOW}. In addition, because choosing to include only one or two relative features in the antecedent would generate a very large number of rules and including more than five relative features in the antecedent would generate very few rules, the minimum number of relative features allowable in the antecedent per rule was fixed to three and the maximum number of antecedents were allowed to be four, which produced an amount of variability in the rules deemed acceptable by the research team (not too many and not too few). Thus, an example rule for course completers could be that 30% of students who had a medium (=2) engagement score on relative features F_1, F_2, and F_7 were likely to persist in the course, while an example rule for course leavers could be that 50% of students who had a medium (=2) engagement score on relative features F_4 and F_5, and low (=1) engagement score on relative features F_6 were likely to drop the course.

TABLE 7 Frequency of top 30 relative features according to learning management system interaction activity type

#	Frequency of top 30 features	
	Data set 1 (31 courses)	**Data set 2 (32 courses)**
1	Quiz submission—6	Quiz submission—5
2	Grades—3	Grades—3
3	Wiki—4	Wiki—2
4	Canvas—5	Canvas—3
5	Attachment—5	Attachment—5
6	Quiz—5	Quiz—2
7	Assignment submission—2	Assignment submission—3
8	Assignment—0	Assignment—4
9	Modules—0	Modules—3
10	Syllabus—0	Syllabus—0
11	Discussion forums—0	Discussion forums—0
12	Announcements—0	Announcements—0

To determine the optimal number of rules to generate, we tested between 20 and 70 rules in increasing increments of five by varying the rules' support and confidence values on which the number of rules ARM generates also depends. We stopped generating rules at 70, as the number of unique rules generated for course completers and course leavers approached saturation as we reached 70 rules, which became the upper bound for the number of rules tested. In other words, generating 80 or 90 rules or an even higher number of rules resulted in unique rules lesser than those obtained from the 70 generated rules. The support of a rule measures how frequently the itemset appears in the data set among all generated rules and the confidence of a rule measures its accuracy, that is, how often the rule is found to be true among the data [1,44]. The range for both support and confidence is between 0 and 1 (or 0% and 100%) and a minimum of 10% threshold is recommended for support values [1,6]. Higher support and confidence values in the algorithm decrease the total number of rules generated. The output confidence values in this study were always 100% because the generated rules were unique to completers or leavers only.

The total number of generated rules, rules after pruning, and unique rules for each case of the desired number of rules are presented in Table 9, along with their respective support values. The generated rules included duplicate rules and rules that were subsets of others. The rules were pruned to remove redundancies as they could introduce bias in the analysis. The rules unique to completers and leavers were determined manually using Microsoft Excel and the duplicate rules were removed. We decided to use all the unique rules in our next step in calculating student engagement. The support values to be entered in the ARM algorithm ranged from 36.2% (for 70 rules) to 39.6% (for 20 rules) for completer-based rules, and 47.4% (for 70 rules) and 51% (for 20 rules) for leaver-based rules. The input confidence values for each case of the desired number of rules ranged from 70% to 95%, these confidence values were manually

FIGURE 2 Association rule mining (ARM) process

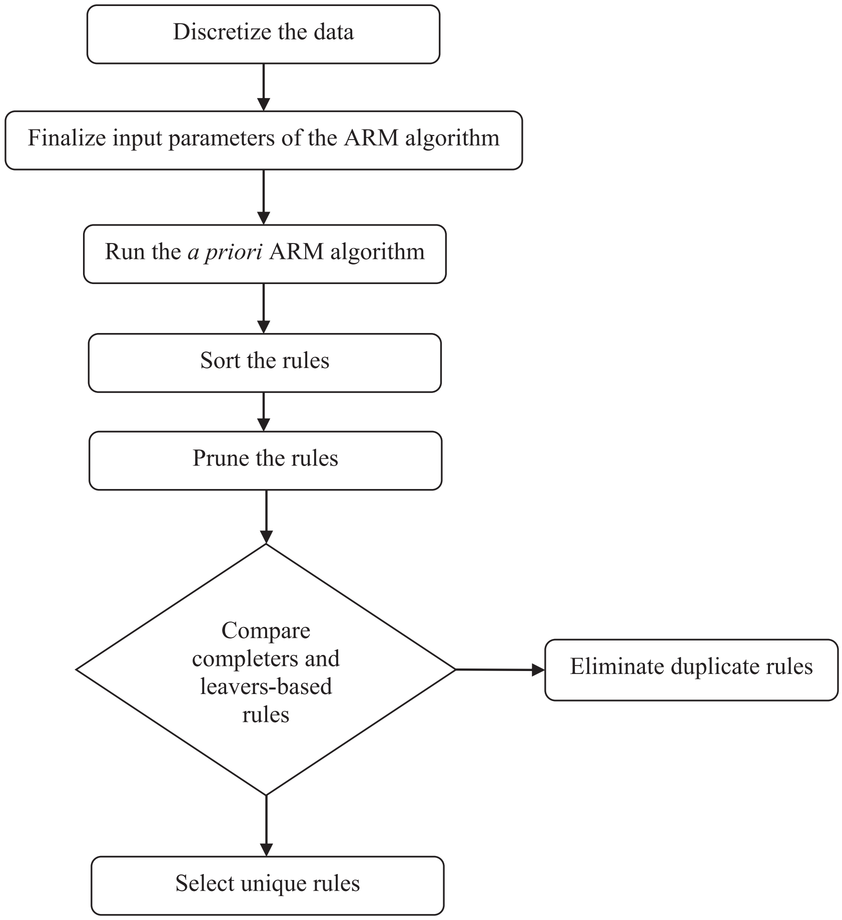

TABLE 8 Association rule mining final problem representation

Student ID	F_1	F_2	F_3	...	F_Z	Persistence
1	1	3	3	...	1	1
2	1	2	1	...	2	0
3	3	2	2	...	3	1
...	...	...	...	...	...	...
N	2	3	1	...	2	1

adjusted to acquire the rules from 20 to 70. Generating 70 rules for both completers and leavers yielded the largest numbers of unique completer-based rules (48) and leaver-based rules (38), respectively. Hence, we chose 70 rules moving forward with the analysis. Supporting Information Appendices A and B show the completer-based unique rules and leaver-based unique rules.

In this study, all the unique rules were used in computing the student engagement scores. However, we note that some ARM researchers select the most important rules using the lift criterion, the ratio of a rule's confidence and support values, and a standard measure for ARM [44,50]. A large lift value is a strong indication that a rule is important and reflects a true connection between consequent and antecedent [50]. In this study, we did not use a lift to select rules as the lift values for all the rules was 1 (100%) implying that all the rules were important.

An example completer-based rule and leaver-based rule generated using the association rule mining process is shown below. These rules were randomly selected (i.e., there was no specific reason or rationale for choosing them). Readers are directed to Table 6 when reading the explanation of the rules below, as this will help better understand and appreciate the definition of the different rules generated in this study.

The following is a completer-based rule:

$\{F3\_quiz.sub2\_3 = MED, \quad F4\_quiz.sub1\_2 = MED, F5\_quiz.sub2 = MED\} \rightarrow \{Persistence = Yes\}$,

where,

- *F3_quiz.sub2_3* represents the difference between the maximum change in the number of quiz submissions for all students in the class and an individual student's change in the number of quiz submissions, across analysis windows 2 and 3,

- *F4_quiz.sub1_2* represents the difference between an individual student's change in the number of quiz submissions and the minimum change in the number of quiz submissions for all students in the class, across analysis windows 1 and 2, and
- *F5_quiz.sub2* represents the difference between the maximum number of quiz submissions by a student in the class and the number of quiz submissions by an individual student, during analysis window 2.

This rule had support of 37.4%, which means that the relative features *F3_quiz.sub2_3 = MED*, *F4_quiz.sub1_2 = MED*, and *F5_quiz.sub2 = MED* appeared together in the data set 37.4% times. All three relative features in this rule had medium levels of student engagement relative to the average student in the course. In other words, medium levels of student engagement relative to the average student in the course on the activities related to quiz submissions are indicative of a student persisting in an online undergraduate engineering course.

Next is a leaver-based rule:

{*F6_quiz3 = LOW*, *F6_wiki3 = LOW*, *F6_assignment.sub3 = LOW*} → {*Persistence = No*},

where,

- F6_quiz3 represents the difference between the time spent on quizzes by an individual student and the minimum time spent on quizzes by a student in the class, during analysis window 3,
- F6_wiki3 represents the difference between the time spent on wiki pages by an individual student and the minimum time spent on wiki pages by a student in the class, during analysis window 3, and
- F6_assignment.sub3 represents the difference between the number of assignment submissions by an individual student and the minimum number of assignment submissions by a student in the class, during analysis window 3.

This rule had support of 47.4%, which means that the relative features *F6_quiz3 = LOW*, *F6_wiki3 = LOW*, and *F6_assignment.sub3 = LOW* appeared together in the data set 47.4% times. All three relative features in this rule had low levels of student engagement relative to the average student in the course. This is to say, the low levels of student engagement relative to the average student in the course on activities related to quizzes, wiki pages and assignment submissions are indicative of a student not persisting in an online undergraduate engineering course.

4.4 | Engagement score determination

The process used in computing student engagement scores using the rules generated by ARM is shown in Figure 3 and more details follow in this section. Using the final set of association rules for leavers and completers, we evaluated eight different candidate student engagement scores, shown in Table 10. The scores were calculated for each student in the testing data set based on the percentage of unique completer-based rules met (*X*) and the percentage of unique leaver-based rules met (*Y*).

TABLE 9 Summary of desired, generated, pruned, and unique completer-based and leaver-based rules

#	Desired number of rules	Completers				Leavers			
		Support (%)	Generated rules	Pruned rules	Unique rules	Support (%)	Generated rules	Pruned rules	Unique rules
1	20	39.6	20	20	12	51.0	19	18	10
2	25	38.8	26	26	17	50.9	25	23	14
3	30	38.5	30	28	17	50.0	31	28	17
4	35	38.0	37	35	24	49.4	36	32	21
5	40	37.8	39	37	25	49.0	41	36	24
6	45	37.7	45	42	30	48.5	48	42	30
7	50	37.2	50	47	35	48.5	48	42	30
8	55	36.9	55	51	38	48.0	55	46	33
9	60	36.7	60	55	40	47.5	61	50	35
10	65	36.5	65	59	43	47.4	69	53	37
11	70	36.2	71	65	48	47.4	69	55	38

Logistic regression was then used to evaluate the efficacy of each candidate student engagement score in predicting students' online course-level persistence. A different candidate engagement score served as the predictor variable in each model and the dependent variable across all models was persistence, with values 0 = leavers and 1 = completers. Notably, we needed to up-sample the data before conducting these analyses to correct the imbalance between course completers and course leavers in the data set [15,40]. There were approximately 12 times as many course completers as there were course leavers in this study. Such a large majority class (one 10 times or larger than the minority class) can introduce bias into logistic regression analysis, which can, in turn, affect the precision and accuracy of predictions about the minority class [32,40]. This imbalance was handled in the statistical software R [39] using the Synthetic Minority Over-sampling Technique (SMOTE), which uses either up-or down-sampling methods to balance unevenly distributed data sets, depending on majority or minority class [15]. SMOTE was used in this study to up-sample the minority class in the data set, the course leavers, by creating synthetic cases [15].

We used eight different SMOTE ratios ranging between 1:1.1 and 1:9, being within the boundaries associated with the actual ratio of leavers to completers (1:12) in the data set, to analyze the stability of each candidate engagement score. In other words, logistic regression was applied to the data for each of eight sampling ratios of leavers to completers and for each candidate engagement score as a predictor of persistence, for a total of 64 models. We favored for the final score those whose effectiveness in predicting course-level persistence remained high. Six performance measures were considered to evaluate the goodness of each engagement score in predicting students' course-level persistence: sensitivity (the proportion of true positives predicted as such), specificity (the detection of true negatives predicted as such), accuracy (the rate of total correct predictions), precision (the rate of correct positive predictions), error rate (the rate of total incorrect predictions), and AUC (area under the ROC curve, which signifies how well the model distinguishes between two classes) [27]. Except for error rate, high values on the other performance metrics indicate greater effectiveness of engagement score.

Table 11 presents the output of these analyses, with the rows representing the eight different candidate engagement scores and the columns, the eight sampling ratios. Notably, every candidate measure of student engagement consistently predicted student persistence to a statistically significant level ($p < .05$) except [inv(X) + inv(Y)] and [inv(Y) − inv(X)].

When a sampling ratio of 1:1.1 was used, most performance metrics (specificity, accuracy rate, and precision rate) were highest for the engagement score [$\log(X)/(\log(X)+\log(Y))$]. When a sampling ratio of 1:3 was used, the highest performance metric in each category was distributed randomly across the eight candidate engagement scores, not suggesting any engagement score to be better than the other. However, for the remaining sampling ratios (1:2, 1:4, 1:5, 1:6, 1:7, and 1:9), most of the performance metrics were highest for the engagement score [$X - Y$]. The performance measures for the engagement score [$X-Y$] for the different sampling ratios of leavers to completers are shown in Table 12. The error rate for [$X-Y$] decreased as the sampling ratio increased from 1:1.1 to 1:9. Every other performance metric except specificity and AUC (fluctuates up and down as sampling ratio increases) increased with increasing sampling ratio from 1:1.1 to 1:9. Specificity which is the accuracy with which true leavers are predicted as such decreases in value is

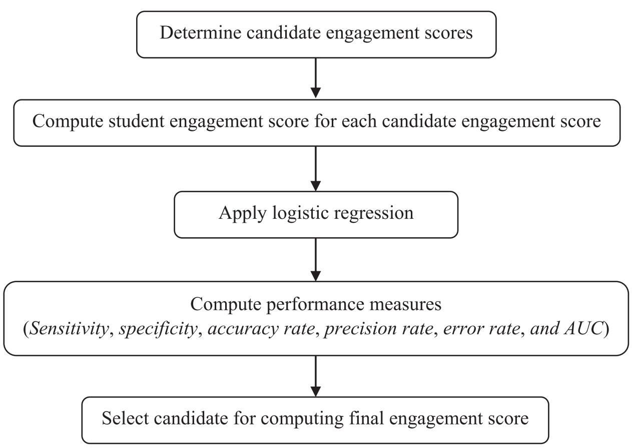

FIGURE 3 Computing student engagement score using the rules

TABLE 10 Different candidate student engagement scores

#	Candidate student engagement scores
1	$X - Y$
2	$\frac{X}{(X+Y)}$
3	$\sqrt{X^2 + Y^2}$
4	$\frac{1}{X} + \frac{1}{Y}$
5	$\frac{1}{Y} - \frac{1}{X}$
6	$\log(X) - \log(Y)$
7	$\log(X) + \log(Y)$
8	$\frac{\log(X)}{\log(X) + \log(Y)}$

TABLE 11 Logistic regression output

Measures of engagement	Logistic regression coefficients							
	1:1.1	1:2	1:3	1:4	1:5	1:6	1:7	1:9
$[X - Y]$	0.04*	0.04*	0.04*	0.04*	0.04*	0.03*	0.04*	0.04*
$[X/(X+Y)]$	3.58*	3.40*	4.17*	3.48*	3.55*	3.37*	3.55*	3.43*
$[\mathrm{sqrt}(X^2+Y^2)]$	−0.01*	−0.01*	−0.01*	−0.01*	−0.01*	−0.01*	−0.01*	−0.01*
$[\mathrm{inv}(X)+\mathrm{inv}(Y)]$	−0.45	−0.69*	−0.75	−0.57	−0.89*	−0.56	−0.63	−0.43
$[\mathrm{inv}(Y)-\mathrm{inv}(X)]$	−0.05	−0.01	0.68	0.83	0.12	0.71	0.31	0.71
$[\log(X)-\log(Y)]$	1.83*	1.75*	1.95*	1.67*	1.79*	1.63*	1.76*	1.69*
$[\log(X)+\log(Y)]$	−0.37*	−0.38*	−0.32*	−0.32*	−0.31*	−0.34*	−0.35*	−0.29*
$[\log(X)/(\log(X)+\log(Y))]$	4.88*	3.03*	5.08*	3.65*	3.06*	3.39*	3.56*	3.32*

Note: Dependent variable – persistence (0 = leavers, 1 = completers).

*$p < .05$.

TABLE 12 Performance measures for the engagement score $[X - Y]$

Performance measures	1:1.1	1:2	1:3	1:4	1:5	1:6	1:7	1:9
Sensitivity	68.12	90.14	93.39	95.44	97.14	97.95	98.75	99.38
Specificity	62.77	42.86	27.38	22.22	19.84	13.10	13.89	13.10
Accuracy rate	65.56	74.38	76.88	80.79	84.26	85.83	88.14	90.75
Precision rate	66.62	75.93	79.42	83.07	85.83	87.12	88.92	91.14
Error rate	34.44	25.62	23.12	19.21	15.74	14.17	11.86	9.25
Area under the curve	70.2	78.7	79.6	78.5	78.8	77.5	79.1	78.3

acceptable as the sampling ratio increases from 1:1.1 to 1:9 because the number of leavers in comparison with the completers decrease. Hence, we selected the candidate $[X - Y]$ for calculating the final student engagement score.

4.5 | Engagement score computation process

Supporting Information Appendix C summarizes the step-by-step approach used in computing the final engagement score. Three subprocesses are outlined: preparing the data required to conduct ARM, conducting ARM, and computing the LMS engagement score using the completer- and leaver-based rules.

5 | DISCUSSION AND INTERPRETATION

In this study, a measure for undergraduate engineering students' engagement in online courses based on data describing their interactions with an online course LMS was introduced. The results from this study suggest that the best engagement score was found to be the mathematical difference between the percentages of unique completer-based rules and leaver-based rules met by each student.

Out of 162 possible relative features, the final set of unique completer-based rules included 29 and the unique leaver-based rules included 19 relative features. The frequency with which these 29 and 19 relative features occurred among the completer- and leaver-based rules is shown in Table 13. The signifier for each relative feature includes its feature type, course activity type, and corresponding analysis windows. For example, relative feature "*F1_quiz.sub3 = MED*" (#10 under the completer-based rules) represents the difference between an individual student's number of quiz submissions and the average number of quiz submissions by all students in the class during the last analysis window (analysis window 3) in the last three analysis windows selected for leavers and the random three analysis windows selected for completers (refer to Table 6 for details related to the numbering and types of relative features). As described previously, the discretization of the data resulted in three levels of student engagement relative to the average

TABLE 13 Frequency of relative features for completer- and leaver-based unique rules

#	Completer-based rules		Leaver-based rules	
	Relative feature	**Frequency**	**Relative feature**	**Frequency**
1	F6_quiz.sub1 = LOW	20	F6_assignment.sub3 = LOW	27
2	F6_quiz1 = LOW	12	F6_quiz.sub3 = LOW	19
3	F1_quiz1 = MED	11	F6_quiz3 = LOW	19
4	F6_quiz.sub2 = LOW	11	F6_assignment.sub1 = LOW	9
5	F6_quiz.sub3 = LOW	11	F6_assignment3 = LOW	5
6	F1_quiz2 = MED	8	F4_assignment.sub2_3 = MED	4
7	F1_quiz3 = MED	7	F6_attach3 = LOW	4
8	F2_quiz1_2 = MED	7	F6_quiz.sub2 = LOW	4
9	F1_quiz.sub1 = MED	6	F3_assignment.sub1_2 = MED	3
10	F1_quiz.sub3 = MED	6	F6_assignment.sub2 = LOW	3
11	F6_quiz2 = LOW	6	F6_canvas3 = LOW	3
12	F1_quiz.sub2 = MED	4	F6_grades3 = LOW	3
13	F2_quiz2_3 = MED	4	F6_wiki3 = LOW	3
14	F6_assignment.sub1 = LOW	4	F6_quiz.sub1 = LOW	2
15	F6_quiz3 = LOW	4	F6_quiz1 = LOW	2
16	F2_quiz.sub1_2 = MED	3	F5_assignment.sub3 = MED	1
17	F6_assignment.sub3 = LOW	3	F6_quiz2 = LOW	1
18	F7_quiz123 = MED	3	F7_grades123 = MED	1
19	F3_assignment.sub2_3 = MED	2	F8_assignment.sub123 = LOW	1
20	F4_assignment.sub1_2 = MED	2		
21	F6_assignment.sub 2 = LOW	2		
22	F1_assignment.sub3 = MED	1		
23	F2_quiz.sub2_3 = MED	1		
24	F3_assignment.sub1_2 = MED	1		
25	F3_quiz.sub2_3 = MED	1		
26	F4_quiz.sub1_2 = MED	1		
27	F5_quiz.sub2 = MED	1		
28	F5_assignment.sub2 = MED	1		
29	F5_assignment.sub1 = MED	1		

student in the course (i.e., 1 = LOW, 2 = MED, 3 = HIGH) and the same can be seen in Table 13. Returning to "*F1_quiz.sub3 = MED*," this rule refers to the difference between an individual student's number of quiz submissions and the average number of quiz submissions by all students in the class during analysis window 3 as being in the "MED" levels of student engagement relative to other students. Overall, most of the relative features that appeared in the final set of unique completer-based rules included the combinations of "LOW" and "MED" levels of student engagement relative to the average student in the course. In the final set of unique leaver-based rules, the relative features that most appeared included "LOW" levels of student engagement relative to the average student in the course, except for a few rules including relative features with "LOW" and "MED" levels of student engagement. This makes sense as students who leave (or plan to leave) the course will be relatively less engaged in the course than other students. In both the unique completer- and leaver-based rules none of the relative features included "HIGH" levels of student engagement relative to the average student in the course. In total, approximately 47% of relative features included "MED" levels of student engagement, 30% included "LOW" levels of student engagement, and 23% included "HIGH" levels of student engagement relative to the average student in the class. As the relative features with "HIGH" levels of student engagement were the least in comparison with "LOW" and "MED" levels of student engagement, they might have not appeared as much in the rules that uniquely distinguish course completers from the course leavers.

Turning our attention to trends across our findings, the sixth relative feature type "*F6*" (see Table 6) refers to the difference between the time spent by an individual student and the minimum time spent by a student in the class in a particular analysis period (refer to Table 13). This relative feature type appeared the most in both unique completer-based and leaver-based rules. Approximately 51% of relative features of the completer-based rules and 91% of relative features of the leaver-based rules were related to feature "*F6*." The 91% dominance of F6 appeared in conjunction with "LOW" levels of student engagement in all the leaver-based unique rules. This particularly makes sense as the relative feature type F6 represents that a student with "LOW" levels of engagement is close to the minimum time spent by a student in the class on a specific activity.

Furthermore, the relative features related to time spent on quizzes, number of quiz submissions, and number of assignment submissions were dominant in both the completer- and the leaver-based rules. This finding echoes the already published work by Crossley et al. [18] and Cohen [17], in which in addition to lecture views, assignment submissions, and number of assessments distinctively predicted course completers. Also, only 4% (approx.) of leaver-based rules additionally included relative features related to time spent on assignments, grades, wiki pages, attachments, and the course canvas site overall.

Approximately 92% of the relative features that appeared in the leaver-based rules (compared to just 44% of the completer-based rules) were associated with LOW levels of student engagement relative to the average student in the course on different activities on the LMS. This finding supports work from Cohen [17], which reported that relatively low measurements on different LMS activities (assignment, course view, discussion forum, and resource view) are indicative of dropout cases. Ten relative features were common to both the unique completer- and leaver-based rules, as shown in Table 14. When the relative features listed in Table 14 are removed from the relative features listed in Table 12, some interesting patterns emerge. Only the relative features with "MED" ratings remain for the completer-based rules and the leaver-based rules are made of mostly relative features with "LOW" ratings. Plus, separately, the leaver-based rules are made of 92% "LOW" ratings whereas half of the completer-based rules do not have all "LOW" ratings and instead have at least one "MED" rating. This implies that students with a combination of "LOW" and "MED" levels of engagement relative to the average student in the course relate to students persisting and completing the course. On the contrary, students with mostly "LOW" levels of engagement relative to the average student in the course relate to students' dropping out from the course.

TABLE 14 Frequency of common relative features for completer- and leaver-based unique rules

#	Relative feature	Frequency of appearance among unique rules: Completers	Leavers
1	F6_quiz.sub1 = LOW	20	2
2	F6_quiz1 = LOW	12	2
3	F6_quiz.sub2 = LOW	11	4
4	F6_quiz.sub3 = LOW	11	19
5	F6_quiz2 = LOW	6	1
6	F6_assignment.sub1 = LOW	4	9
7	F6_quiz3 = LOW	4	19
8	F6_assignment.sub3 = LOW	3	27
9	F6_assignment.sub2 = LOW	2	3
10	F3_assignment.sub1_2 = MED	1	3

6 | IMPLICATIONS AND FUTURE WORK

In this section, we present the benefits/implications of this study to researchers interested in the educational data mining space. In addition, we provide potential directions for future work related to this study.

6.1 | Benefits

This study includes several potential benefits and the same are presented in this section. The analysis described provides researchers in the educational data mining space a new approach to conduct their own investigations related to online student engagement, an important construct to the study of student persistence in online courses. As no one correct approach to calculating a student engagement score exists, we recommend researchers carefully explore and modify the approach to their data in addition to applying our method of evaluating the best online student engagement score to other data sets. Long-term potential implications of such measures include helping online course instructors identify students at risk of dropping a course.

6.2 | Future work

The work presented in this paper describes a novel approach to numerically represent student engagement among online undergraduate engineering students using their LMS interaction data. Random forest was used to select the relative features used in association rule mining to generate completer-based rules and leaver-based rules. A total of 48 unique completer-based rules and 38 unique leaver-based rules were obtained. The best student engagement score using these rules was found to be the mathematical difference between the percentage of completer rules and the percentage of leaver rules met by each student. The relative features did not include "HIGH" levels of engagement relative to the average student in the course in both the unique completer- and leaver-based rules, investigating this further could be a potential direction for future work. The relative feature type *F6* appeared the most in both unique completer- and leaver-based rules and the data in this study is not sufficient to provide a rationale for this, hence this needs to be explored further. The next steps for this study include combining the student–LMS interaction data and student attribute data to predict students' persistence decisions, that is, we would like to test whether a model that includes both student–LMS interaction data and student attribute data is superior to a model that uses just either of these in the prediction of students' course-level persistence intentions.

ACKNOWLEDGMENTS

This paper is based on research supported by the National Science Foundation (NSF) under Award Number 1825732. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Javeed Kittur http://orcid.org/0000-0001-6132-7304

REFERENCES

1. R. Agrawal, T. Imieliński & A. Swami Mining association rules between sets of items in large databases. Proc. 1993 ACM SIGMOD Internat. Conf. Manag. Data, Washington, DC, 1993, pp. 207–216.
2. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, *Fast discovery of association rules*, Advances in Knowledge Discovery and Data Mining (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds.), 1st ed., vol. **12**, AAAI, Menlo Park, CA, 1996, pp. 307–328.
3. R. Agrawal & R. Srikant Fast algorithms for mining association rules. Proc. 20th int. Conf. Very Large Databases, VLDB, vol. 1215, Madison, 1994, pp. 487–499.
4. E. Aguiar, N. V. Chawla, J. Brockman, G. A. Ambrose & V. Goodrich Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention. Proc. Fourth Internat. Conf. Learn. Anal. Knowl., Indianapolis, 2014, pp. 103–112.
5. I. E. Allen, J. Seaman, R. Poulin & T. T. Straut Online report card. Tracking online education in the United States, Babson Survey Res. Group and Quahog Res. Group, Wellesley, 2016.
6. B. Azarnoush, J. M. Bekki, B. L. Bernstein & G. C. Runger An associative based approach to analyzing an online learning environment, 120th ASEE Annu. Conf. Expo., Atlanta City, 2013.
7. G. Balakrishnan and D. Coetzee, *Predicting student retention in massive open online courses using hidden Markov models*, Electr. Eng. Comput. Sci. Univ. Calif. Berkeley **53** (2013), 57–58.
8. H. Baytiyeh and M. Naja, *Identifying the challenging factors in the transition from colleges of engineering to employment*, Eur. J. Eng. Educ. **37** (2012), no. 1, 3–14.
9. J. Bohan and N. Stack, *Do weekly online homework assignments promote student engagement and enhance learning in first year psychology students?* Psychol. Educ. Rev. **38** (2014), no. 2
10. M. L. Bote-Lorenzo & E. Gómez-Sánchez Predicting the decrease of engagement indicators in a MOOC. Proc. Seventh Internat. Learn. Anal. Knowl. Conf., 2017, pp. 143–147.
11. J. Bowers and P. Kumar, *Students' perceptions of teaching and social presence: A comparative analysis of face-to-face and online learning environments*, Int. J. Web-Based Learn. Teach. Technol. (IJWLTT) **10** (2015), no. 1, 27–44.
12. S. Boyer & K. Veeramachaneni Transfer learning for predictive models in massive open online courses. Internat. Conf. Artif. Intel. Education, Springer, Cham, Switzerland, 2015, pp. 54–63.
13. S. Brunhaver, J. Bekki, E. Lee & J. Kittur Understanding the factors contributing to persistence among undergraduate engineering students in online courses. Companion Proc. 9th Internat. Conf. Learning Anal. Knowl., Tempe, 2019.
14. T. E. Chatman Jr., Understanding the relationship between student engagement and persistence among community college students, Doctoral dissertations and projects, Liberty University, Lynchburg, VA, 2020.
15. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique*, J. Artif. Intell. Res. **16** (2002), 321–357.
16. H. Coates, *A model of online and general campus-based student engagement*, Assess. Eval. Higher Educ. **32** (2007), no. 2, 121–141.
17. A. Cohen, *Analysis of student activity in web-supported courses as a tool for predicting dropout*, Educ. Technol. Res. Dev. **65** (2017), no. 5, 1285–1304.

18. S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara & R. S. Baker Combining click-stream data with NLP tools to better understand MOOC completion. Proc. Sixth Internat. Conf. Learn. Anal. Knowl., Edinburgh, 2016, 6-14.
19. M. D. Dixson, *Measuring student engagement in the online course: The Online Student Engagement Scale (OSE)*, Online Learn. **19** (2015), no. 4.
20. M. Fei & D. Y. Yeung Temporal models for predicting student dropout in massive open online courses. 2015 IEEE Internat. Conf. Data Mining Workshop (ICDMW), Atlantic City, 2015, pp. 256–263.
21. G. Gercek, N. Saleem, and D. Steel, *Implementing cloud based virtual computer network labs for online education: Experiences from a phased approach*, Int. J. Online Biomed. Eng. (iJOE) **12** (2016), no. 03, 70–76.
22. P. Gregori, V. Martínez, and J. J. Moyano-Fernández, *Basic actions to reduce dropout rates in distance learning*, Eval. Progr. Plan. **66** (2018), 48–52.
23. S. Halawa, D. Greene, and J. Mitchell, *Dropout prediction in MOOCs using learner activity features*, Proc. Sec. Eur. MOOC Stakeholder Summit **37** (2014), no. 1, 58–65.
24. C. Hart, *Factors associated with student persistence in an online program of study: A review of the literature*, J. Interact. Online Learn. **11** (2012), no. 1, 42.
25. C. R. Henrie, R. Bodily, R. Larsen, and C. R. Graham, *Exploring the potential of LMS log data as a proxy measure of student engagement*, J. Comput. Higher Educ. **30** (2018), no. 2, 344–362.
26. J. F. Humber Student engagement in online courses: A grounded theory case study, Doctoral dissertation, Univ. Alabama Libraries, Tuscaloosa, 2018.
27. G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning, (vol. **112**) Springer, New York, NY, 2013, p. 18.
28. A. Karaksha, G. Grant, S. Anoopkumar-Dukie, S. N. Nirthanan, and A. K. Davey, *Student engagement in pharmacology courses using online learning tools*, Am. J. Pharm. Educ. **77** (2013), no. 6, 125.
29. J. Kittur, J. Bekki & S. Brunhaver Learner Analytics in Engineering Education: A detailed account of practices used in the cleaning and manipulation of learning management system data from online undergraduate engineering courses. Proc. Am. Soc. Eng. Educ., 2020.
30. R. F. Kizilcec, C. Piech & E. Schneider Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. Proc. Third Internat. Conf. Learn. Anal. Knowl., Leuven, 2013, pp. 170–179.
31. S. Kocdar, A. Bozkurt, and T. Goru Dogan, *Engineering through distance education in the time of the fourth industrial revolution: Reflections from three decades of peer-reviewed studies*, Comput. Appl. Eng. Educ. **29** (2021), no. 4, 931–949.
32. T. J. Lakshmi & C. S. R. Prasad A study on classifying imbalanced datasets. 2014 First IEEE Internat. Conf. Netw. Soft Comput. (ICNSC2014), Guntur, 2014, pp. 141–145.
33. B. Liu, W. Hsu & Y. Ma Integrating classification and association rule mining, KDD, 98 (1998), pp. 80–86.
34. N. Mohamad, N. B. Ahmad, and S. Sulaiman, *Data preprocessing: a case study in predicting student's retention in MOOC*, J. Fundam. Appl. Sci. **9** (2017), no. 4S, 598–613.
35. P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, and C. Delgado Kloos, *Re-defining, analyzing and predicting persistence using student events in online learning*, Appl. Sci. **10** (2020), no. 5, 1722.
36. L. V. Morris and C. L. Finnegan, *Best practices in predicting and encouraging student persistence and achievement online*, J. Coll. Stud. Retent.: Res.Theory Pract. **10** (2008), no. 1, 55–64.
37. T. Petty and A. Farinde, *Investigating student engagement in an online mathematics course through windows into teaching and learning*, J. Online Learn. Teach. **9** (2013), no. 2, 261–270.
38. S. J. Quaye, S. R. Harper, and S. L. Pendakur (eds.), Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations. Routledge, 2019.
39. R Core Team, R: A language and environment for statistical computing, R Found. Stat. Comput., Vienna, Austria, 2017, https://www.R-project.org/
40. A. H. A. Rahim, N. A. Rashid, A. Nayan & A. R. Ahmad Smote approach to imbalanced dataset in logistic regression analysis. Proc. Third Int. Conf. Comput. Math. Stat. (iCMS2017), Springer, Singapore, 2019, pp. 429–433.
41. J. C. Richardson and T. Newby, *The role of students' cognitive engagement in online learning*, Am. J. Distance Educ. **20** (2006), no. 1, 23–37.
42. C. Romero and S. Ventura, *Educational data mining: a review of the state of the art*, IEEE Trans. Sys. Man Cybernetics, Part C (Appl. Rev.) **40** (2010), no. 6, 601–618.
43. S. G. Salvo, K. Shelton, and B. Welch, *African American males learning online: Promoting academic achievement in higher education*, Online Learn. **23** (2019), no. 1, 22–36.
44. J. E. Seaman, I. E. Allen, and J. Seaman, Grade increase: Tracking distance education in the United States, Babson Survey Res. Group, 2018.
45. P. Shea and T. Bidjerano, *A national study of differences between online and classroom-only community college students in time to first associate degree attainment, transfer, and dropout*, Online Learn. **20** (2016), no. 3, 14–15.
46. B. E. Shelton, J. L. Hung, and P. R. Lowenthal, *Predicting student success by modeling student interaction in asynchronous online courses*, Distance Educ. **38** (2017), no. 1, 59–69.
47. O. Sneed Fostering student engagement through an online community of learning: A mixed methods action research dissertation, Doctoral dissertation, Arizona State Univ., Tempe, 2019.
48. M. Stewart, T. Stott, and A. M. Nuttall, *Student engagement patterns over the duration of level 1 and level 3 geography modules: Influences on student attendance, performance and use of online resources*, J. Geogr. Higher Educ. **35** (2011), no. 01, 47–65.
49. Y. Sun, A. K. Wong & Y. Wang An overview of associative classifiers. DMIN, Las Vegas, 2006, pp. 138–143.
50. P. N. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, Pearson Educ., India, 2016.
51. C. Taylor Stopout prediction in massive open online courses, Doctoral dissertation, Mass, Inst. Technol., Cambridge, 2014.
52. A. Trumbore, *Rules of engagement: Strategies to increase online engagement at scale*, Change: Mag. Higher Learn. **46** (2014), no. 4, 38–45.
53. J. M. Vytasek, A. Patzak, and P. H. Winne, *Analytics for student engagement*, Machine learning paradigms, Springer, Cham, Switzerland, 2020, pp. 23–48.

54. J. Watts, *Assessing an online student orientation: Impacts on retention, satisfaction, and student learning*, Tech. Commun. Q. **28** (2019), no. 3, 254–270.
55. A. F. Wise, J. M. Vytasek, S. Hausknecht, and Y. Zhao, *Developing learning analytics design knowledge in the" Middle Space": The student tuning model and align design framework for learning analytics use*, Online Learn. **20** (2016), no. 2, 155–182.
56. S. Young and M. A. Bruce, *Classroom community and student engagement in online courses*, J. Online Learn. Teach. **7** (2011), no. 2, 219–230.
57. X. Zeng, C. Yu, Y. Liu, X. Hu, Q. Hao, Y. Jiang, W. Dai, K. Zhang, and B. Teng, *The construction and online/offline blended learning of small private online courses of principles of chemical engineering*, Comput. Appl. Eng. Educ. **26** (2018), no. 5, 1519–1526.

AUTHOR BIOGRAPHIES

Javeed Kittur is a PhD Candidate in the Engineering Education Systems and Design program at Arizona State University. He received his bachelor's and master's degree in Electrical Engineering from BVB College of Engineering and Technology and The National Institute of Engineering, India. His research interests include data analytics in engineering education, persistence of students in online undergraduate engineering, and engineering faculty preparedness in cognitive, affective, and psychomotor domains of learning.

Jennifer Bekki is an Associate Professor in The Polytechnic School within the Ira A. Fulton Schools of Engineering at Arizona State University. She received her MS and PhD degrees in Industrial Engineering from Arizona State University. Her research aims to: (1) increase our understanding of the factors contributing to student persistence decisions among the populations of online students and doctoral students in STEM and (2) identify opportunities for systemic changes that address inequities arising from racism and sexism within STEM graduate education.

Samantha Brunhaver is an assistant professor in The Polytechnic School of the Ira A. Fulton Schools of Engineering at Arizona State University. She received her MS and PhD degrees in Mechanical Engineering from Stanford University. Her research interests include engineering education and workforce development, academic and professional persistence, and online learning.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: J. Kittur, J. Bekki, and S. Brunhaver, Development of a student engagement score for online undergraduate engineering courses using learning management system interaction data, Comput. Appl. Eng. Educ. 2022; 30:661–677. https://doi.org/10.1002/cae.22479