ASEE'S VIRTUAL CONFERENCE
At Home with Engineering Education
JUNE 22 - 26, 2020 #ASEEVC
Paper ID #29804

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

Mr. Javeed Kittur, Arizona State University

Javeed Kittur is a doctoral student (Engineering Education Systems & Design) at Arizona State University, USA. He received a Bachelor's degree in Electrical and Electronics Engineering and Master's in Power System from India in 2011 and 2014 respectively. He has worked with Tata Consultancy Services as Assistant Systems Engineer from 2011-2012, Bangalore, India. He has worked as an Assistant Professor (2014 to 2018) in the department of Electrical and Electronics Engineering, KLE Technological University, India. He is a certified IUCEE International Engineering Educator. He is awarded with the 'Ing.Paed.IGIP' title at ICTIEE, 2018.

Dr. Jennifer M Bekki, Arizona State University

She teaches courses in the engineering and manufacturing engineering programs as well as programs in the Engineering Education Systems and Design PhD program. Her research interests include topics related to student persistence, STEM doctoral student experiences, faculty mentorship and development, modeling and analysis of complex manufacturing systems, and the development of new discrete event simulation methodologies. Bekki is the co-director of the interdisciplinary, National Science Foundation supported CareerWISE research program, which strives to: 1) understand the experiences of diverse women who are pursuing and leaving doctoral programs in science and engineering and 2) increase women's persistence in science and engineering doctoral programs through the development and dissemination of an online resilience and interpersonal communication training program.

Dr. Samantha Ruth Brunhaver, Arizona State University, Polytechnic campus

Samantha Brunhaver is an Assistant Professor of Engineering in the Fulton Schools of Engineering Polytechnic School. Dr. Brunhaver recently joined Arizona State after completing her M.S. and Ph.D. in Mechanical Engineering at Stanford University. She also has a B.S. in Mechanical Engineering from Northeastern University. Dr. Brunhaver's research examines the career decision-making and professional identity formation of engineering students, alumni, and practicing engineers. She also conducts studies of new engineering pedagogy that help to improve student engagement and understanding.

Learner Analytics in Engineering Education: A Detailed Account of Practices Used in Cleaning and Manipulating Learning Management System Data from Online Undergraduate Engineering Courses

Abstract

This is a research paper that provides a concrete example for other engineering education researchers of how Learning Management System (LMS) interaction data from online undergraduate engineering courses can be prepared for analysis. We provide the rationale and details involved in choices related to data preparation, feature creation, and feature selection as part of a larger National Science Foundation-funded study dedicated to developing a theoretical model for online undergraduate engineering student persistence. LMS interaction data provides details about students' navigations to and submissions of different course elements including quizzes, assignments, discussion forums, wiki pages, attachments, modules, the syllabus, the gradebook, and course announcements. The sample dataset presented here includes 32 courses from three ABET accredited fully online engineering degree programs, electrical engineering, engineering management, and software engineering, offered at a large, public, southwestern university. The analysis demonstrated in this paper will ultimately be combined with associative classification to discover relationships between student-LMS interactions and persistence decisions.

Introduction

Online education is growing in acceptance and is rapidly increasing in use, including in higher education as evident by increasing enrollments over the last decade (25.9%-2012, 27.1%-2013, 28.3%-2014, 29.7%-2015 and 31.6%-2016) [1]. It provides better flexibility, ease of access to students and relatively lower costs in comparison to in-person face-to-face courses. Additionally, with limited physical infrastructure and facility requirements, online education provides an opportunity to offer courses with large student enrollments [2]. Many institutions are now offering online courses, and the academic leaders of these institutions underscore that learning through online courses will be a critical strategy for their institutions to be successful in the long run [1].

Despite the benefits that online education offers, online courses face challenges with respect to student attrition; specifically, attribution in online courses is higher than in face-to-face courses [3-4]. One study reported that the attrition rate of students in online courses is 10-15% higher than that for the in-person face-to-face courses [5]. In another study on distance education, it was reported that the dropout rates in online courses are approximately three times higher than the in-person course [6]. In yet another study, the attrition rates in online education were reported to be 10-20% higher than that for the face-to-face courses [7]. Finally, in a study by Mishra [8], the attrition rate in online courses from 27 open universities was reported to be 15.265, and another study intended to mitigate the attrition rates in online graduate program, it was found that the attrition rates were in between 28% to 48% [9].

With the concern about the attrition rates of students in online courses, many researchers have tried to understand the associated reasons. Huitt [10], for example, reports that motivation is an internal state of an individual, which is closely associated with engagement and is responsible in guiding one's behavior in an online setting. Several other studies showed that if students have higher motivation to complete a course, they are less likely to drop the course [11], [12], [13], [14]. Hart [15] also identified different factors relating to the persistence of students in online courses such as sense of belonging in the community, motivation, time management skills, communication with instructor, online learning satisfaction, and peer and family support. Shelton, Hung and Lowenthal [16] found that the frequency of social interactions in online courses is an important factor in understanding the persistence and success of students. Finally, in another study, it was reported that the parameters that influence the successful completion of online courses include prior academic achievement, financial assistance, continuous academic enrollment and previous information technology training [17].

Some educational researchers have also developed models using the data from online courses to predict student performance and outcomes [18]. Aguiar et. al, [19], for example, used the electronic portfolios of freshmen engineering students to predict persistence. In another study, Morris & Finnegan [20] use students' academic and demographic variables to predict students' course-level persistence decisions. Submitting assignments, solving exercises and watching lectures were used to predict engagement in MOOCs [21], and another study to predict the dropouts in MOOCs used features related to learner activity such as assessment performance, assignment skips, and video skips and lags [22].

In educational research, engagement is an important parameter that correlates with persistence [23]. Student engagement can be operationalized in different ways, but generally include components related to time spent on activities [23-24]. For online students, the online course delivery format generates a rich, temporal stream of data describing how individual learners interact with the online course, and previous work [25-26], shows the promise of utilizing LMS data to model engagement. Work by Macfadyen and Dawson [27], for example, uses the student-LMS interactions data to predict the students' academic achievement using regression modeling. In Bovo, et al. [28], the authors use the LMS interactions to predict performance of students in online courses, and they use machine learning algorithms to cluster students using features created from the LMS data. In a study conducted to understand the MOOC completion, the student-LMS interactions and messages posted on discussion forum were considered [29]. And in yet another study [30], based on the students' LMS interaction logs, the students were classified to understand the relationship between these classified groups and their performance. Bosch and co-authors [31] studied the interactions of underrepresented students in online STEM courses using features created from the student-LMS interactions. Finally, within the field engineering education specifically, [32], the LMS interactions of freshmen engineering students was analyzed through learning analytics to understand students' course engagement and performance [32], and work by Castro et. al [33], uses LMS interactions to understand the behavior of online engineering master's students.

While there is clearly an abundance of literature related to conducting research using online LMS interaction data within higher education, and some within the engineering education research community, the intensity of focus within engineering education is decidedly less than within in the general higher education community. One of the reasons for this could be that engineering education researchers in this space are not confident about how to get started with related research topics (another possible reason could be that the researchers do not consider the LMS interaction data to be a string indicator of student engagement). Compounding this is that most available literature utilizing use LMS interaction to build models related to student outcomes provide information related to the background of the study, participants, data collection, methodology and algorithms. However, specifics about data cleaning, manipulation and data preparation are not detailed enough to help novice educational researchers. In this paper, our goal is to provide a detailed account of the practices we used in cleaning and manipulating LMS data from online undergraduate engineering courses at a large, public southwestern university. We specifically operationalize engagement as a function of interactions with the online LMS and ultimately plan to use these engagement patterns to distinguish between students who persist and those who drop. Correspondingly, the analysis includes the creation of features that describe the difference between a student's engagement behavior and the "typical" behavior across all students in their same online course. This notion of examining differences from the social norm is in line with previous research by Coates [34] who identified four categories of student engagement, each of which are a function of both social and academic norms.

In the treatment of the literature, we are intentionally broad in what we are considering as online learning, including both MOOCS, fully online, and hybrid courses in support of a particular degree program. Our specific research context is focused on fully online courses in support of ABET accredited engineering programs, but we find relevance in the research of other researchers of online education and choose to draw from that literature in providing background context.

The subsequent sections in this paper discusses the details related to the sample dataset used in this study, then we provide a detailed account of the steps followed in data preparation, feature creation and feature selection. The work presented here is part of a larger National Science Foundation-funded study [35] dedicated to developing a theoretical model for online undergraduate engineering student persistence based on student LMS interaction activities and patterns.

Data Set

Any study utilizing LMS interaction data requires researchers to actually have access to the associated data. At the university where this research is situated, a separate university organization oversees delivery of all the university's online courses. This organization's charge also includes a research mission. Correspondingly, they support related faculty research projects by providing access to the LMS interaction data, and the process of acquiring the data for this study included building a relationship between our research team and this entity on campus, pitching our research ideas to them, and having them subsequently agree to provide access to the data.

In this section, we describe the dataset that will serve as a testbed to illustrate our process for data preparation and cleaning, feature creation, etc. The sample dataset includes 994,439 rows of students' activity data from 1,725 students enrolled in 32 different online engineering courses offered at a large, public, southwest university during the fall 2018 and spring 2019 semesters. Of the 32 courses, 15 are from software engineering, 14 are from engineering management, and 3 are from electrical engineering. Table 1 shows the corresponding enrollment by degree program among these courses. Fig. 1 shows the percentage of students within each degree program who persisted to course completion and who chose to drop the course before completing it. The overall percentage of students who dropped the course at some point during the semester in these 32 courses is 10.09%. All 32 courses utilized the LMS platform Canvas, and all courses were 7.5 weeks in duration. Of note is that there is a discrepancy in the number of online courses offered between Fall 2018 and Spring 2019. This project focuses only on data obtained from Canvas, and during the period of this data collection, the university at which this data was collected was migrating from one LMS platform to Canvas. Faculty had the 2018–2019 academic year to transition from the previous LMS to canvas, and more hand done so by Spring 2019 than Fall 2018.

Table 1 Student enrollment in fall 2018 and spring 2019

	Electrical Engineering	**Engineering Management**	**Software Engineering**
Fall 2018	92	85	-
Spring 2019	190	352	1006

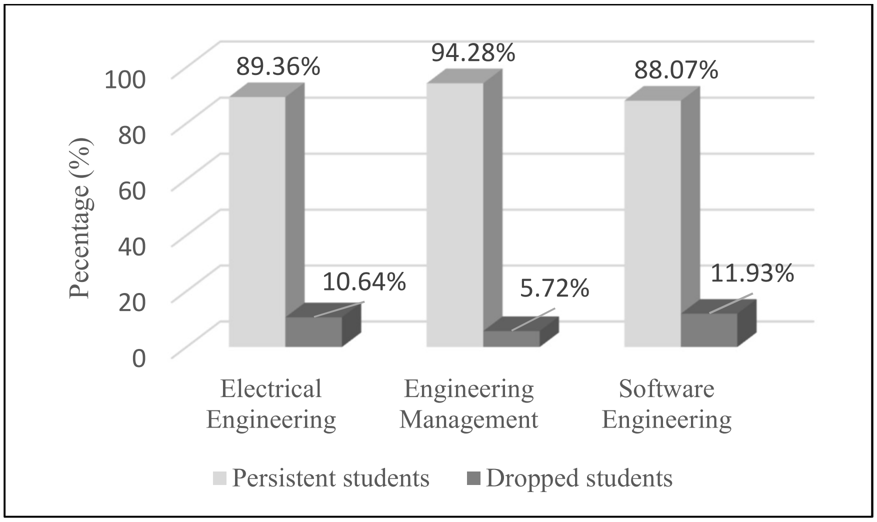


Fig. 1 Percentage persistent and dropped students across three degree programs

Within Canvas, students engage in different activities such as browsing course content, monitoring grades, taking part in discussions with peers and instructors, and taking assessments. The raw data contains detailed logs of how each student navigates through different sections/parts in the LMS related to a course. A sample of the structure of the raw data is shown in Table 2 with de-identified student_id and course_id. The different columns in the dataset are event time, student id, course id, event type, action, object type, object name, event details, and enrollment status among others. Student id is unique for each student and course id is unique for each course offered in a semester. The last column in Table 2 with the enrollment status is useful in identifying whether a student has persisted or withdrawn the course. The LMS interaction data provides details about students' navigations to and submissions of different course elements including quizzes, assignments, discussion forums, wiki pages, attachments, modules, syllabus, grades, and course announcements.

Table 2. Structure of the raw data

eventtime	student_id	course_id	eventtype	action	object_name	enrl_status
10/10/2018 9:21:33	A	2018Fall B_X	NavigationEvent	NavigatedTo	quizzes:quiz	ENRL
10/15/2018 9:22:18	A	2018Fall B_X	NavigationEvent	NavigatedTo	attachment	ENRL
10/11/2018 19:54:17	B	2018Fall B_X	NavigationEvent	NavigatedTo	syllabus	ENRL
10/16/2018 15:55:03	B	2018Fall B_X	AssessmentEvent	Submitted	-	ENRL
10/22/2018 10:06:53	C	2018Fall B_X	NavigationEvent	NavigatedTo	modules	ENRL
10/22/2018 17:11:47	C	2018Fall B_X	NavigationEvent	NavigatedTo	grades	ENRL
10/13/2018 23:05:59	D	2018Fall B_X	AssignableEvent	Submitted	-	WDRW
10/16/2018 23:45:24	E	2018Fall B_X	Event	Modified	-	WDRW
10/24/2018 0:00:55	F	2018Fall B_X	NavigationEvent	NavigatedTo	announcements	WDRW

In this study, a student in a course offering belongs to one of the following three groups:

1. Students who enrolled in the course but dropped it during the typical drop period at the beginning of a semester (i.e., they never really participated in the course).
2. Students who enrolled in the course but ultimately withdrew (i.e., they paid for and participated in the course, but chose to drop the course before completing it).
3. Students who enrolled and remained enrolled until the course was complete.

Data Preparation

In data preparation, the first step is data cleaning. Specifically, activity data logs that take place before the course start date (e.g., perusing the course site or syllabus) or after the course completion date (e.g., checking final grades) are removed. Additionally, for this study, information related to course grade were removed, as they are not within the purview of this study. Finally, we chose to exclude the first three days of course activity for each course, with the idea that many students will be dropping for reasons not relevant to the study at hand. The statistical software R was used to do the data cleaning. Readers who are interested in learning more about the application of R to social science research are directed to tidyverse, which provides a curated collection of R packages designed for data science.

Once the initial data cleaning is complete, the next step is to determine the amount of time that students spent on different activities within their LMS. These calculations are typically called ‘time-on-task’ [36]. The time spent on an activity by a student in a course is calculated by taking the difference between two consecutive clicks made by the student on the LMS. In this particular dataset, to do so, the date and time were split into two different columns, and the time difference between the two rows was assigned as the activity time associated with the activity.

Utilizing LMS interaction data then typically requires an additional step related to time-on-task calculations. The LMS for online courses do not typically log students out automatically. Correspondingly, there are often lags between consecutive clicks (and corresponding time-on-task calculations) that are extremely long (e,g, 12 - 24 hours, or more). In practice, it is difficult to believe that a student was working on any one activity for such a long period; instead, most researchers who use LMS data assume that these long periods of time correspond to periods in which the learners stepped away from their LMS to engage in different activities. Correspondingly, a cap is typically placed on activity times, but there is broad disagreement between what this cap should be. Some studies have used 30 minutes as this cap, others have 10 minutes, 15 minutes, 20 minutes, or one hour [37-42].

The activity cap should naturally be a function of the duration of expected activities [36]. In this dataset, raw activity data ranged from less than 1 minute to more than 1000 minutes. It is not possible to know which activity lengths are genuine and which are artificial, but we felt that an activity cap of less than 90 minutes would potentially be removing legitimate content evaluation activities. Three different activity caps were considered: no cap, 90 minutes, and 180 minutes. Based on exploration of the data itself, including consideration of outliers, the research team selected 90 minutes as the cap. We acknowledge that there is not one right answer for what the activity cap should be and encourage researchers in this area to carefully consider what makes sense for their own context. Correspondingly, once the times-on-task were calculated, if the activity time was found to be greater than 90 minutes, then it was replaced with 90 minutes. Additionally, the final activity time of each student in a course was also assigned to be 90 minutes.

This study aims at identifying students who are at risk and who will eventually drop the course by analyzing the student LMS interaction. An assumption made in this study is that the LMS interaction patterns are different for students who persist the course than the students who drop the course and that we need interaction data from a sufficiently long amount of time in order to calculate features related to behavior that could be used to detect such differences. We selected a period of three days as the basic unit of analysis for feature creation, and we called these three-day periods a “window”. The three-day window allowed us to recognize that students may choose different days and times to work on different tasks (which would not be recognizable for daily analyses), but also gave us enough data to detect temporal patters (vs. bucketing student’s interactions weekly in the 7.5 week courses, which we believed could gloss over important details).

After we removed the first three days of course activity (for reasons described at the beginning of this section), the data set included a total of 16 three-day windows, as shown in Fig 2. We would ultimately use these windows to calculate time-on-tasks (i.e., total time on the LMS in window 1 or total time checking site wiki (or content) pages during window 3). Correspondingly, when considering what time is part of what window, it is necessary to consider the carryover time from the previous activity at the beginning of the window and limit the excess time at the end of the window. For example, if an activity started at 11:50pm on Day 6 of the course and ended at 12:18am on Day 7 of the course, part of that activity belongs to Window 1, and part to Window 2. Before proceeding with the carryover times and excess times calculations, the data was segregated and labelled as “previous day”, “day one”, “day two”, and “day three” as related to each window. An example of this labeling for two windows is given in Table 3.

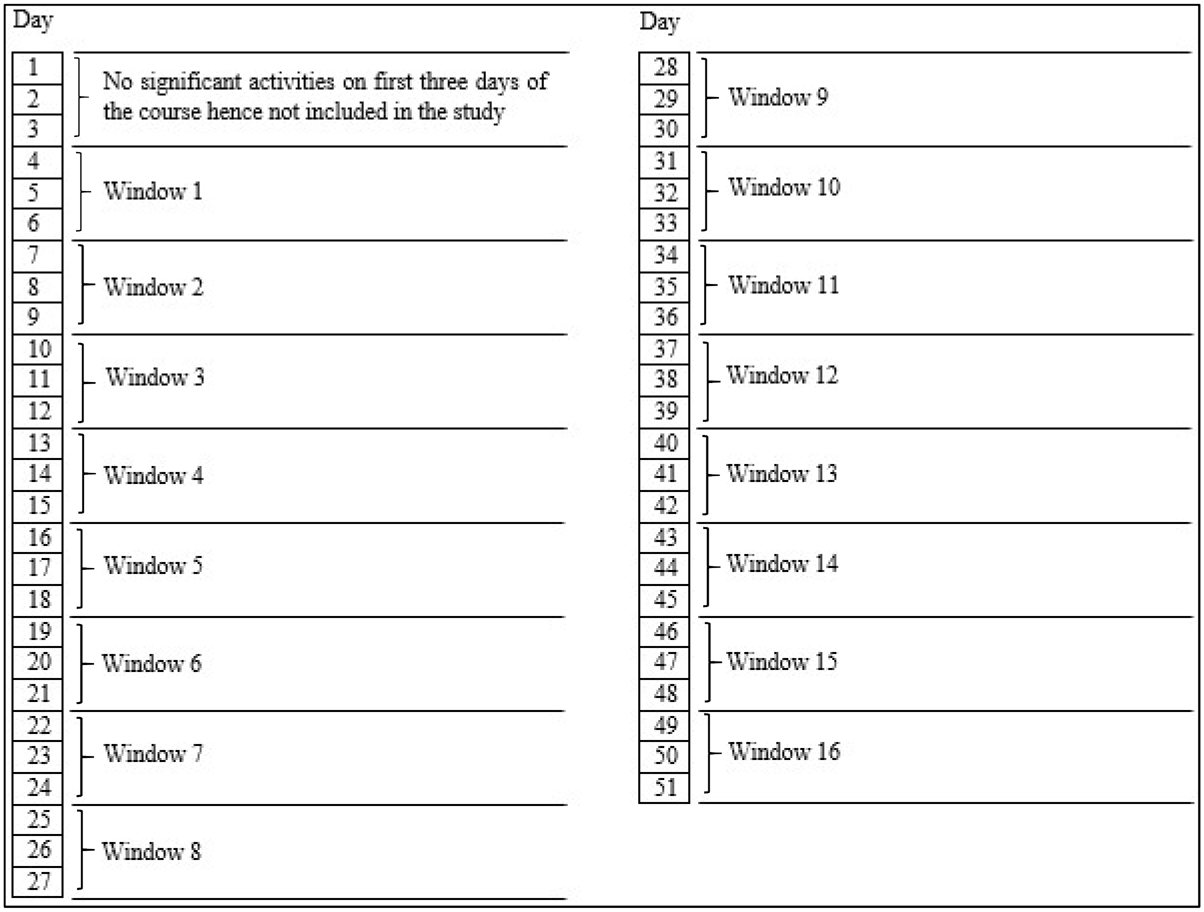


Fig. 2 Segregation of 7.5 weeks in 16 windows

The carryover times were calculated using the following steps; (1) identify the day prior (previous day) to the start of the window, (2) find the last click time of each student for the day prior to start of the window, (3) to implement the 90-minute cap on activity time, if the time is greater than 10:30:00 PM, subtract this time from midnight – we refer to this time as X minutes, (4) subtract X from the activity time, which we refer to as Y minutes, (5) assign Y mins to the activity for the day prior to the start of the considered window, and include that as a new entry in the window. To limit the excess times, the following steps were used: (1) find the last click time of each student for the last day of the window (day three) (2) to implement the 90-minute cap on activity time, if the time is greater than 10:30:00 PM, subtract this time from midnight – we refer to this as X minutes, (3) assign X minutes to the activity pertaining to that click. In both carryover times and excess times calculations, the time was checked if it is greater than 10:30:00 PM because if it was less than 10:30:00 PM, then activity will have been curved by the 90 minutes threshold and will not continue the next day. At the conclusion of this step, the data is organized into 16 windows, and time-on-task for every activity within each window is calculated.

Table 3 Example of days assigned to calculate excess and carryover times

	Window 1 **(10/10, 10/11 & 10/12)**	**Window 2** **(10/13, 10/14 & 10/15)**
Previous day	10/09/2018	10/12/2018
Day one	10/10/2018	10/13/2018
Day two	10/11/2018	10/14/2018
Day three	10/12/2018	10/15/2018

The next step involved categorizing these times-on-task into different types of activities for each student within each window. Many previous researchers have discussed features that are relevant for working with LMS-interaction data. We build on this literature, and the final LMS activity categories, including associated references, that we utilize in this study are presented in Table 4.

Table 4. Activity types calculated for each student in each window in each class

No.	**Activity Type**	**Reference**
1	Time spent on quizzes	[21]
2	Time spent on assignments	[31], [43]
3	Time spent on discussion forum	[43], [44]
4	Time spent on wiki pages	[43]
5	Time spent on attachments	[31]
6	Time spent on modules	[44]
7	Time spent on syllabus	[44]
8	Time spent on grades	[44]
9	Time spent on announcements	[44]
10	Time spent on canvas	[31], [34], [43]
11	Number of quiz submissions	[31], [43], [44]
12	Number of assignment submissions	[31], [43]

The cleaned data was used to calculate the time spent on each different type of activity. To do so, the first step was to isolate the individual time-on-task (or frequency data) for each student and each activity type within each window, and then to sum up these values for each student and activity type within each window. Students who did not have activity of a particular type during a particular window were assigned a corresponding value of 0. At the conclusion of this step, the data for each class was organized into columns that corresponded to activity types within each window. The rows corresponded to students, and the values contained within each column corresponded to the amount of time (or frequency) spent by that student on the corresponding activity type within that window.

Before finalizing the data preparation to yield categorized time-on-task data for each student in each window, by class, we considered the significance of outliers. This was of particular relevance to our study because we are considering the relationship between each students' activity and "the typical" activity for the course. Outliers would disproportionately impact what was considered "typical" within a particular class. Correspondingly, across all students enrolled in a class, for each activity type (e.g., "quiz submissions") within each window, we used quartiles and the inter-quartile range (IQR) to calculate outliers. Specifically, building on other published work [45], we considered any activity time greater than $Q_3 + 1.5*IQR$ to be an outlier. The average value of the feature for a specific activity was calculated excluding the outliers, and this average value was used to replace the outliers. Following the replacement of outliers, the data preparation was complete and had a format as shown in Table 5. The column 2 to 9 in Table 5 show the time spent on different activities in the first window period by 10 students from an electrical engineering course.

Table 5. Structure of the data at the end of data preparation phase

student	tquiz	tassignment	tdforum	twiki	tattach	tmodules	tsyllabus	tgrades	Status
A	57.36	0.422	0.383	278.5	193.1	111.9	4.31	3.8	ENRL
B	15.01	0.266	0	30	54.43	0	0.46	0	ENRL
C	18.81	0.1	2.45	239.7	291.1	138.2	0.01	0.18	ENRL
D	9.96	0.16	1.58	0	91.13	0.76	0.01	0.55	ENRL
E	48.68	0.85	1.01	184.8	32.03	1.41	0	0.52	ENRL
F	93	0	0.23	5.58	27.88	90.08	2.36	0	ENRL
G	9.58	4.13	0.57	92.50	88.91	61.75	3.35	0.28	WDRW
H	2.73	0.1	0.06	1.46	6.5	0.23	2.30	0	WDRW
I	109.8	0.42	0.57	227.8	16.95	183.02	0	0.52	WDRW
J	0	0	2.13	0.03	94.6	1.21	0.01	0	WDRW

All time-on-task calculations were done using R. The same calculations were also done "manually" in excel, and several test cases were used to verify if the approach used in calculating the features in R matches with the more manual (and so more easily verifiable) calculations done in excel.

Feature creation

Following data preparation, the next step in conducting the analysis on LMS interaction data is to calculate features derived from the times-on-task (within each window, by activity type) that can be used to detect differences between persisting and dropping students. For example, Fig 3, shows one example of the total time spent on LMS for one dropped and one persistent student from electrical engineering course.

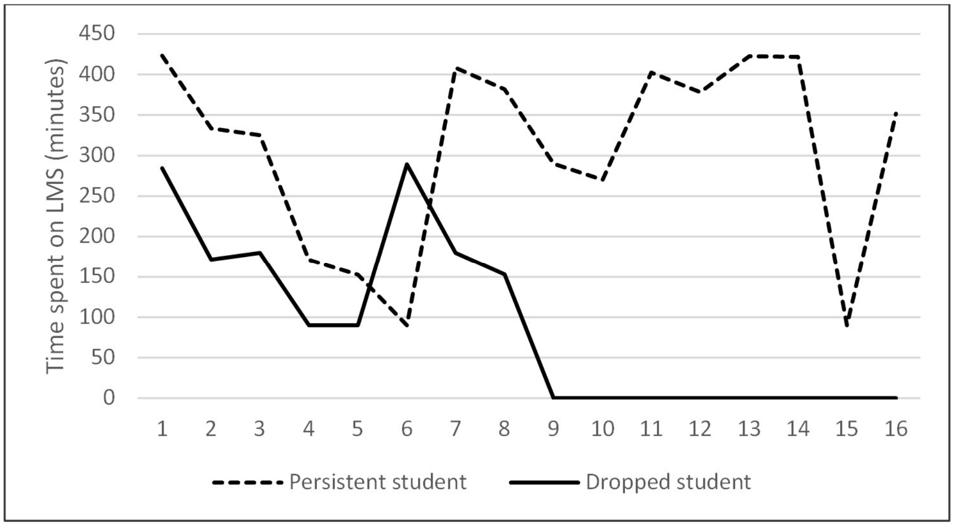


Fig 3. Example of total time on canvas for one dropped and one persistent student

As described previously, we are specifically interested in creating features that both describe the differences between student's engagement and the overall class engagement in a course and also capture the temporal nature of engagement. For dropped students, we assume that these features would be most distinctive in the period just before a student drops, while for persisting students, we assume their behavior to be nondistinctive across the duration of the course. To generate these relative and temporal features, we need to determine the number of windows on which to base our analysis such that we have enough data prior to a student dropping that we could use it to potentially discriminate their behavior from a "typical" behavior of a student who persists. In the selection of data to be used for each student, we would select the corresponding number of windows prior to their drop date as the analysis window for a dropped student, and a randomly selected equal length amount of time for students who persist. We would then calculate features based on this data and perform feature selection to determine which features are most discriminatory between the groups.

To assist with the determination of how many windows of activity we will consider in our analysis, we considered the number of students who dropped during each window of the course. This is relevant because students who dropped the course before accruing enough time in the course to meet our analysis window would not be eligible for consideration in the study. Given that the total percentage of dropped students is so small in comparison with persistent students, we were careful not to select an analysis length that was still long enough to capture relevant behavior, but not so long that it significantly reduced the number of dropping students who remain within our dataset.

Table 6 shows the percentage of students who dropped, by window. Based on Table 6, we decided to use three windows of data as our analysis period and removed from the dataset any students who dropped in windows 1, 2 and 3 (because they were not enrolled long enough to meet our analysis window requirements). Also, the choice of 3 windows provided us the minimum amount of time in order to be able to capture the variance. Correspondingly, our final data set included

1667 students, of which 121 students did not persist (i.e., final percentage drop out of students = 7.25%).

Table 6 Percentage loss of data (of dropped students)

Window	Number of students dropped	Percentage (%)	Cumulative (%)
1	19	11.58	11.58
2	18	10.97	22.56
3	12	7.31	29.87
4	8	4.87	34.75
5	14	8.53	43.29
6	21	12.8	56.09
7	5	3.04	59.14
8	7	4.26	63.41
9	8	4.87	68.29
10	11	6.70	75.00
11	10	6.09	81.09
12	3	1.82	82.92
13	6	3.65	86.58
14	1	0.60	87.19
15	8	4.87	92.07
16	5	3.04	95.12
Last day	4	2.43	97.56
After last day	4	2.43	100.0

For each course, the time-on-task (or frequency) values shown in Table 5, were used for each student to calculate a number of relative (i.e., comparing their engagement and temporal activity with that of other students in class) features within the three-window analysis timeframe. Specifically, for each of the time-on-task (or frequency) categories shown in Table 5, we calculated each of the features shown in Table 7. With this, the total number of features that were generated and used for analysis were 216.

Table 7: Notation and Features

Notation

T_{ijk} – Time spent by student i in course j in unit time k

n_{jk} – Number of students in course j in unit time k

M_{ijk} – Number of submissions by student i in course j in unit time k

D_{ij} – Duration of the course considered for a student i and course j

N – number of windows

Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time.

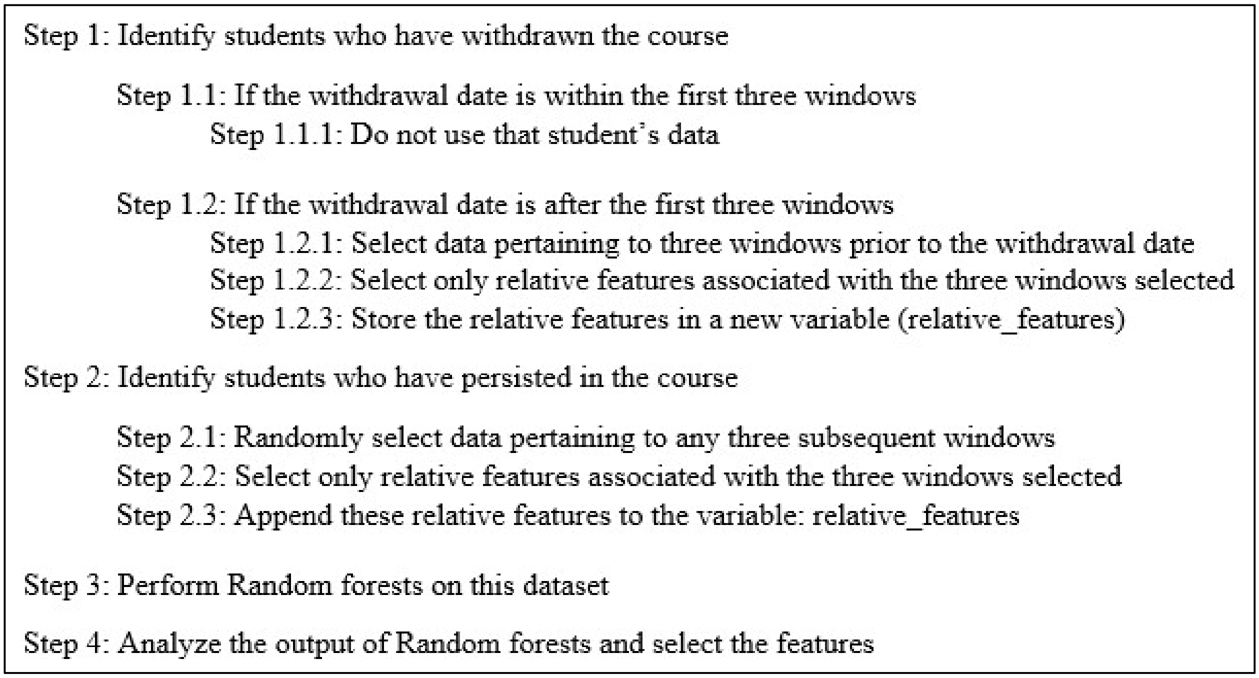
$$T_{ijk} - \frac{\sum_{i \in n_{jk}} T_{ijk}}{n_{jk}} \quad \forall\, k \in D_{ij}$$

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 given unit of time.

$\left(T_{ijk} - T_{ijk'}\right) - \left[\frac{\sum_{i \in n_{jk}} \left(T_{ijk} - T_{ijk'}\right)}{n_{jk}}\right]$ $\forall\, k, k' \in\ D_{ij}\ and\ k < k'$
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 given unit of time $max_{i \in j}\left(T_{ijk} - T_{ijk'}\right) -\ \left(T_{ijk} - T_{ijk'}\right)$ $\forall\, k, k' \in\ D_{ij}\ and\ k < k'$
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 given unit of time $\left(T_{ijk} - T_{ijk'}\right) -\ min_{i \in j}\left(T_{ijk} - T_{ijk'}\right)$ $\forall\, k, k' \in\ D_{ij}\ and\ k < k'$
Difference between the maximum time spent by a student in the class and the time spent by an individual student, in a given unit of time. $max_{i \in j,k}(T_{ijk}) - T_{ijk} \quad \forall\, i \in j, k\ and\ k \in\ D_{ij}$
Difference between the time spent by an individual student and the minimum time spent by a student in the class, in a given unit of time $T_{ijk} - min_{i \in j}(T_{ijk}) \quad \forall\, i \in j, k\ and\ k \in\ D_{ij}$
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 $\frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2 - \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2}{n_{jk}}$ $\forall\, i \in j, k\ and\ N \in 3\ and\ k \in D_{ij}$
Difference between the maximum variance of the time spent by a student in the class and the variance of time spent by an individual student across three different windows $max\left\{\frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2\right\} - \frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2$ $\forall\, i \in j, k\ and\ N \in 3\ and\ k \in D_{ij}$
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 three different windows $\frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2 - min\left\{\frac{1}{N-1}\sum_{k=1}^{N}\left[T_{ijk} - \frac{\sum_{i=1}^{n_{jk}} T_{ijk}}{n_{jk}}\right]^2\right\}$ $\forall\, i \in j, k\ and\ N \in 3\ and\ k \in D_{ij}$

Feature Selection

The larger objective of this research study is to use the calculated features and appropriately classify the students who persisted and dropped the course. In order to do so, we need to be able to identify which features are most useful in making such distinctions, and the process of doing so is called feature selection. In this study, we use the feature selection that is part of the Random Forest algorithm [46] and use R to conduct this analysis. An overview of the algorithm used for selecting the data for analysis and the subsequent feature analysis can be found in Figure 4.



```
Step 1: Identify students who have withdrawn the course

        Step 1.1: If the withdrawal date is within the first three windows
                Step 1.1.1: Do not use that student's data

        Step 1.2: If the withdrawal date is after the first three windows
                Step 1.2.1: Select data pertaining to three windows prior to the withdrawal date
                Step 1.2.2: Select only relative features associated with the three windows selected
                Step 1.2.3: Store the relative features in a new variable (relative_features)

Step 2: Identify students who have persisted in the course

        Step 2.1: Randomly select data pertaining to any three subsequent windows
        Step 2.2: Select only relative features associated with the three windows selected
        Step 2.3: Append these relative features to the variable: relative_features

Step 3: Perform Random forests on this dataset

Step 4: Analyze the output of Random forests and select the features
```

Fig. 4 Algorithm for analysis data selection and feature selection

We very briefly present the results here for reference, recognizing that the purpose of this paper is not in sharing the results of our research project, but more in documenting our process for use by other researchers in similar work. The results from feature selection using the random forest algorithm are presented in Tables 8 and 9. Table 8 shows the frequency of top 30 features (out of 216 features) that appeared from the output of random forest algorithm. The features that appeared the most in these top 30 features are ranked in order as shown in Table 8. The features related to quiz submissions, grades and wiki pages appeared six times each in the top 30 features. The features related to canvas appeared four times, attachment appeared three times, quiz and assignment submissions appeared twice, assignment appeared once respectively. There were no features related to modules, syllabus, discussion forum and announcements in top 30 features. Hence, the features that can be selected from the output of random forest are features related quiz submission, grades, wiki pages, canvas and attachments.

Table 9 includes the top 10 features that resulted from the feature selection using Random Forest. Out of the top 10 features listed in Table 9, features related to quiz submissions appeared three times, features related to grades and canvas appeared twice and features related to wiki pages, number of assignment submissions and quizzes appeared once.

Table 8. Frequency of top 30 features

Sl. No	Frequency of top 30 features
1	Quiz submission – 6
2	Grades – 6
3	Wiki pages – 6
4	Canvas – 4
5	Attachment – 3
6	Quiz – 2
7	Assignment submission – 2
8	Assignment – 1
9	Modules – 0
10	Syllabus – 0
11	Discussion forum – 0
12	Announcements – 0

Table 9. Top 10 features

#	Description
1	Difference between the time spent by an individual student and the minimum time spent by a student in the class, in a given unit of time
2	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 given unit of time
3	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 given unit of time
4	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
5	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
6	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
7	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
8	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
9	Difference between an individual student's time spent and the average time spent for all students in the class, in a given unit of time
10	Difference between the maximum time spent by a student in the class and the time spent by an individual student, in a given unit of time

Conclusion and Future Work

This study is a part of larger National Science Foundation funded study. This work provides a detailed account of the practices used in data cleaning and manipulation of the learning management system of undergraduate engineering courses. This paper provides the rationale and details involved in choices related to data cleaning, manipulation, and feature creation.

In the sample dataset considered in this study, only 10.09% of the total students represent the dropped students and hence the dataset is clearly imbalanced between persisted and dropped students. In these cases of imbalanced data, it is more difficult to develop models that accurately discriminate between the two groups. Hence, future work will consider the use of a resampling strategy to increase the representation within the dataset of students who dropped. The Synthetic Minority Over-Sampling Technique (SMOTE) [47], is a popular data sampling method which uses both up-sampling and down-sampling strategies depending on the class. SMOTE creates synthetic cases for a minority class by randomly selecting the nearest neighbors. Once we are satisfied with the dataset itself, the features selected from the random forest output will be ultimately combined with associative classification to discover relationships between student-LMS interactions and persistence decisions.

Acknowledgements

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.

References

1. Seaman, J. E., Allen, I. E., & Seaman, J. (2018). Grade Increase: Tracking Distance Education in the United States. *Babson Survey Research Group*.
2. Rovai, A. P., & Downey, J. R. (2010). Why some distance education programs fail while others succeed in a global environment. *The Internet and Higher Education*, *13*(3), 141-147.
3. Frydenberg, J. (2007). Persistence in university continuing education online classes. *The international review of research in open and distributed Learning*, *8*(3).
4. Heyman, E. (2010). *Overcoming student retention issues in higher education online programs: A Delphi study*. University of Phoenix.
5. Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. *Chronicle of higher education*, *46*(23).
6. Brady, L. (2001). Fault lines in the terrain of distance education. *Computers and Composition*, *18*(4), 347-358.
7. Angelino, L. M., Williams, F. K., & Natvig, D. (2007). Strategies to engage online students and reduce attrition rates. *Journal of Educators Online*, *4*(2), n2.
8. Mishra, S. (2017). Open universities in the Commonwealth: At a glance.
9. Oregon, E., McCoy, L., & Carmon-Johnson, L. (2018). Case Analysis: Exploring the Application of Using Rich Media Technologies and Social Presence to Decrease Attrition in an Online Graduate Program. *Journal of Educators Online*, *15*(2), n2.
10. Huitt, W. (2001). Motivation to learn: An overview. *Educational psychology interactive*, *12*.
11. Chen, K-C, & Jang, S-J. (2010). Motivation in online learning: Testing a model of self-determination theory. *Computers in Human Behavior*, 26, 741-752.
12. Hartnett, M., George, A. S., & Dron, J. (2011). Examining motivation in online distance learning environments: Complex, multifaceted and situation-dependent. *The International Review of Research in Open and Distance Learning*, 12(6), 20-38.

13. Muilenburg, L. Y., & Berge, Z. L. (2005). Student barriers to online learning: A factor analytic study. *Distance Education, 26*(1), 29-48.
14. Bekele, T.A. (2010). Motvation and satisfaction in internet-supported learning environments: A review. *Educational Technology & Society*, 13(2), 116-127.
15. Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning*, *11*(1).
16. Shelton, B. E., Hung, J. L., & Lowenthal, P. R. (2017). Predicting student success by modeling student interaction in asynchronous online courses. *Distance Education*, *38*(1), 59-69.
17. Salvo, S. G., Shelton, K., & Welch, B. (2019). African American Males Learning Online: Promoting Academic Achievement in Higher Education. *Online Learning*, *23*(1), 22-36.
18. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.
19. Aguiar, E., Ambrose, G. A. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2014). Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. *Journal of Learning Analytics*, *1*(3), 7-33.
20. Morris, L.B., & Finnegan, C.L. (2008). Best practices in predicting and encouraging student persistence and achievement online. *Journal of College Student Retention*, 10(1), 55-64.
21. Bote-Lorenzo, M. L., & Gómez-Sánchez, E. (2017, March). Predicting the decrease of engagement indicators in a MOOC. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 143-147). ACM.
22. Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the second European MOOC stakeholder summit*, *37*(1), 58-65.
23. Quaye, S. J., & Harper, S. R, (Eds.). (2014). *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*. Routledge.
24. Kuh, G. D., Kinzie, J., Schuh, J. H., & Whitt, E. J. (2011). *Student success in college: Creating conditions that matter*. John Wiley & Sons.
25. Azarnoush, B., Bekki, J. M., Runger, G. C., Bernstein, B. L., & Atkinson, R. K. (2013a). Toward a framework for learner segmentation. *Journal of Educational Data Mining*, 5(2), 102-126.
26. Azarnoush, B., Bekki, J.M., Bernstein, B.L., & Runger, G.C. (2013b) An associative based approach to analyzing an online learning environment, in *Proceedings of the 2013 American Society of Engineering Education (ASEE) Annual Conference,* paper ID #7142.
27. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, *54*(2), 588-599.
28. Bovo, A., Sanchez, S., Héguy, O., & Duthen, Y. (2013, September). Clustering moodle data as a tool for profiling students. In *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE)* (pp. 121-126). IEEE.
29. Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016, April). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 6-14).
30. Akçapınar, G. (2015). Profiling students' approaches to learning through moodle logs. In *Multidisciplinary Academic Conference on Education, Teaching and Learning (MAC-ETL 2015)*.

31. Bosch, N., Crues, R. W., Henricks, G. M., Perry, M., Angrave, L., Shaik, N., & Anderson, C. J. (2018). Modeling key differences in underrepresented students' interactions with an online STEM course. In *Proceedings of the Technology, Mind, and Society* (pp. 1-6).
32. Brozina, C., Knight, D. B., Kinoshita, T., & Johri, A. (2019). Engaged to Succeed: Understanding First-Year Engineering Students' Course Engagement and Performance Through Analytics. *IEEE Access*, *7*, 163686-163699.
33. Castro, M., Menacho, A., & Perez-Molina, C. (2018, April). Mining LMS students' data on online task-based master degree studies. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 661-668). IEEE.
34. Coates, H. (2007). A model of online and general campus-based student engagement. *Assessment & Evaluation in Higher Education*, *32*(2), 121-141.
35. Brunhaver, S., Bekki, J., Lee, E., & Kittur, J. (2019, March). Understanding the factors contributing to persistence among undergraduate engineering students in online courses. In *International Conference on Learning Analytics & Knowledge*.
36. Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015, March). Penetrating the black box of time-on-task estimation. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 184-193). ACM.
37. Chitraa, V., Davamani, D., & Selvdoss, A. (2010). A survey on preprocessing methods for web usage data. *arXiv preprint arXiv:1004.1257*.
38. Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs journal on computing*, *15*(2), 171-190.
39. Ba-Omar, H., Petrounias, I., & Anwar, F. (2007, July). A framework for using web usage mining to personalize e-learning. In *Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)* (pp. 937-938). IEEE.
40. Munk, M., & Drlik, M. (2011, June). Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. In *International Conference on Digital Information and Communication Technology and Its Applications* (pp. 60-74). Springer, Berlin, Heidelberg.
41. Nguyen, Q., Huptych, M., & Rienties, B. (2018, March). Linking students' timing of engagement to learning design and academic performance. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 141-150). ACM.
42. Del Valle, R., & Duffy, T. M. (2009). Online learning: Learner characteristics and their approaches to managing learning. *Instructional Science*, *37*(2), 129-149.
43. Boyer, S., & Veeramachaneni, K. (2015, June). Transfer learning for predictive models in massive open online courses. In *International conference on artificial intelligence in education* (pp. 54-63). Springer, Cham.
44. Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 256-263). IEEE.
45. Su, X., & Tsai, C. L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 261-268.
46. Tan, P. N. (2018). Introduction to data mining (pp. 533).
47. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.