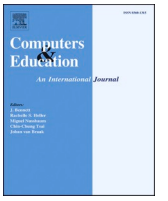
Computers & Education 223 (2024) 105154

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu



Analysis and discovery of procrastination patterns in a language learning MOOC

Jorge A. Vázquez Mendoza [a,*], Cèsar Ferri Ramírez [b], Carlos Monserrat Aranda [b]

[a] Graduate University Master's Degree in Information Management, Universitat Politècnica de València, Spain
[b] Department of Computer Systems and Computation, Universitat Politècnica de València, Spain

ARTICLE INFO

Keywords:
Process mining
Massive open online courses
Procrastination
Language learning

ABSTRACT

Online learning has been gaining a broad notoriety in society in the last years. The ease with which users from all over the world can learn is one of its main attractions. MOOCs are one of these technologies, which have enabled users to learn almost any subject of their interest. The use of MOOCs generates a massive amount of data that has been used by researchers with different goals: to predict dropout, predict grades and determine learning styles. Users of MOOCs face several challenges, some of which have been extensively studied. However, there are some of them, such as procrastination, whose study as a determining factor in the failure of such courses has not been addressed in sufficient depth. Through this paper, the influence of procrastination on the failure rates of users in a MOOC has been uncovered. Furthermore, by using process mining, this research has revealed the existence of a pattern of procrastination and the type of material used during study sessions by the users who fail the course. Speclifically, diverse forms of procrastination have been identified, resulting in differing effects on the educational outcomes attained by the user. These findings possess considerable implications, as they underscore the potential worth of detecting specific procrastination patterns to ascertain users who necessitate supplementary support during their educational pursuits within MOOCs.

1. Introduction

This paper examines the impact of procrastination on users performance in Massive Open Online Courses (MOOCs). We apply data science techniques (process mining) that extract knowledge from log data. Concretely, process mining tools are employed over user behaviour generated by a MOOC on learning Spanish. The analysis of the extracted knowledge reveals patterns associated with procrastination. These insights offer valuable hints for enhancements in course effectiveness and the design of suitable support strategies for users.

The success of MOOCs has revolutionised education across various fields, including language learning. MOOCs are characterised by being able to host numerous users who only need to have access to the internet to follow a course (Aljaraideh, 2019). In 2019, more than 120 million users were enrolled in MOOCs globally, whereas by 2021, this number had risen to around 220 million. Considering just the MIT, in 2023, 3,5 million users were enrolled in their MOOCs (Cagiltay, Toker, & Cagiltay, 2023). This signifies a growth percentage of approximately 80%, underscoring a substantial expansión in the intervening years (Shah, 2021), probably impulsed by Covid-19.

* Corresponding author.
E-mail addresses: jorvazme@alumni.upv.es (J.A. Vázquez Mendoza), cferri@dsic.upv.es (C.F. Ramírez), cmonserr@dsic.upv.es (C.M. Aranda).

https://doi.org/10.1016/j.compedu.2024.105154
Received 30 January 2023; Received in revised form 2 September 2024; Accepted 6 September 2024
Available online 12 September 2024

To motivate users, MOOCs commonly include readings, videos, interactive activities, and gamification (Jarnac de Freitas & Mira da Silva, 2023). Besides, in some domains, such as language learning, these courses contain other types of content. Specifically, in MOOCs related to language learning, the inclusion of tasks that motivate exploration, reflection, production and dialogue among the users is recommended (Teixeira & Mota, 2014). This objective can be met by including audio and pronunciation activities (Read, 2014). Unfortunately, the dropout rate is usually high in MOOCs (Borrella, Caballero-Caballero, & Ponce-Cueto, 2019; Mrhar, Douimi, & Abik, 2021). Generally, users of online courses face a recurring challenge of procrastination (Yao, Sahebi, & Feyzi Behnagh, 2020; Huang, Zhang, Burtch, Li, & Chen, 2021). Besides, these behaviours have shown a strong correlation with the achievement of goals in MOOCs (Diver & Martinez, 2015).

Platforms hosting MOOCs generate considerable data including user demographics, course content browsing records, etc. As the enrolment of users in MOOCs continues to expand, the resultant surge in data production becomes increasingly substantial. This growing influx of data makes manual or semi-automated management and monitoring of user behaviour unfeasible. Here is where several studies have proposed the analysis of this substantial amount of data generated by MOOCs (Doss, Krishnan, Karuppasamy, & Sam, 2024; Liu, Tang, Ouyang, Long, & Liu, 2024; Norman & Nordin, 2024). Henceforth, adopting data analysis techniques, such as process mining, is highly recommended.

Process mining can be defined as a family of techniques based on the areas of data science and process management that provide tools to support the analysis of operational processes based on event logs. Process mining aims to turn event data into insights and knowledge. In general, process mining allows organisations to identify behaviour patterns based on the data generated by their systems. In the educational field, it has been used to model user behaviour and use these models to classify them (Thiede, Fuerstenau, & Barquet, 2018). The emerging field of educational data mining that tries to better understand the educational process through analysing log data gathered specifically from educational environments has been defined as Educational Process Mining (EPM) (Bogarín, Cerezo, & Romero, 2018).

Following this line of reasoning, we have discovered the behavioural patterns of different types of MOOC users (classified according to their use and success in the course) by using the behavioural analysis tool process mining and its relationship with the level of procrastination. For the experimental evaluation, we consider data from a MOOC language course offered by the *Universitat Politècnica de València* (Spain). The course used as a case study is entitled "Basic Spanish 1" in its first edition in 2020. As has been mentioned, the objective of the study is to identify counterproductive behaviour patterns of users (mainly procrastination patterns) and relate this behaviour with failure and/or dropout using process mining. Through the pattern extraction analysis, we have identified that users who show procrastination behaviour within their study sessions fail and dropout of the course. This result is also closely linked to the fact that these users do not properly manage their learning time and do not take advantage of the material available in the course, as do the users who complete it. Hence, we can conclude that detecting procrastination patterns early can significantly help launch mitigation actions, increasing the success ratio of MOOC users.

This paper is organised as follows. In Section 2, we describe basic concepts related to MOOCs. Section 3 introduces process mining and summarises how this technique has been applied in the learning domain. Section 4 presents the MOOC used in the experiments. Section 5 includes the methodology used in the experiments, while Section 6 analyses the results obtained. Section 7 discusses the study's constraints. Finally, Section 8 discusses the conclusions that were drawn and provides further analysis of the results that were obtained.

2. Massive Open Online Courses (MOOCs)

Globally, the use of technology for learning has become increasingly common in society. Among the various technological tools that have been developed for this purpose are open online courses. MOOCs are commonly made up of videos, readings and a certain number of tests and assignments (Mesquita, Toda, & Brancher, 2014).

These courses store and generate a huge amount of user data (demographic information, assessment results, event logs, etc.). This data is useful for researchers and data scientists who seek to improve and/or optimise the learning process of users in a MOOC.

Users in a MOOC do not have a tutor to monitor their learning process as in a traditional class. Therefore, the absence of a tutor implies a set of barriers, among which are: little understanding of the language of the course, lack of an instant feedback and the poor organisation of the user's time (Aljaraideh, 2019). Therefore, the results obtained by the users of the course depend on their ability to self-direct their learning that could be defined as the ability of the individuals to guide their own learning process (Song & Hill, 2007).

In the context of MOOCs, studies to date have focused on four (not necessarily disjoint) aspects of the learning process in such courses:

- **Dropout**: some studies indicate that MOOC dropout rates range from 16% to 48%, based on a study of 5 different courses (Borrella et al., 2019). Broader studies suggest that it's around 90% of MOOC enrollees who do not complete them (Mrhar et al., 2021). In any case, the dropout rate of MOOCs is alarmingly high. Multiple factors have been identified as reasons for user dropout in MOOCs, including time constraints, technical difficulties, course difficulty level, lack of motivation, and additional certification costs, among others. (Wang, Zhao, Wu, & Goh, 2022)
- **Self-regulation of learning**: this is the essential characteristic of self-directed learning. The ability to learn autonomously is directly linked to success in an online course, and this challenge is related to the users' discipline in maintaining a study pace (Maldonado et al., 2016).
- **Interest and/or procrastination**: MOOCs are designed to be self-paced, devoid of fixed schedules, often prompting users to defer tasks until the eleventh hour. Consequently, such procrastination can trigger a decline in motivation, giving rise to difficulties in

sustaining concentration and adhering to the course trajectory (Mutawa, 2023). At times, loss of interest and procrastination are closely related. However, even when the topic of study is of interest to users, they tend to postpone the learning process when they encounter challenges they consider more engaging.

- **Participation**: MOOCs provide forums as spaces for idea exchange, acting as key conduits for interaction between users and instructors. However, there are scenarios where these forums do not always serve as efficient platforms for meaningful stakeholder engagement. Moderated forums, while maintaining decorum, can limit user-to-user interactions. Conversely, when forums are open for free expression, they might not always furnish the necessary support that users require. (Onah, Sinclair, & Boyatt, 2014)

The present work is focused on the search for procrastination patterns of MOOC users based on process mining techniques, trying to relate these patterns with the users' dropout or failure.

3. Process mining and online learning

Process mining is a discipline that, based on the event logs generated by information systems, can extract knowledge from them. This knowledge can be used to discover, monitor or improve the processes of the analysed information systems (Nieves, Ramírez-Quintana, Monserrat, Ferri, & Hernández-Orallo, 2020; Vázquez Mendoza, 2021). Fig. 1 shows the common flow of an analysis using process mining. From databases of events and actions performed by all the system users, the first task is to extract the sequence of events/actions developed by each user. The result of this first task will be as many sequences as users. Process mining algorithms can obtain a graph summarising the extracted activity sequences from this set of sequences. In other words, this graph generalises the sequence of actions that users follow to perform the modelled task. In this research, process models are interaction graphs between the elements analysed within a study session. Process models are characterised by their ability to visualise the analysed behaviour based on the intensity of the colours of each node and the thickness of each edge.

An adapted version of the Process Mining Methodology (PM2) (Van Eck, Lu, Leemans, & Van Der Aalst, 2015) has been used to answer the research questions detailed in section 5. This adapted version of the methodology has seven steps: 1) Definition of objectives, 2) Basic related concepts, 3) Data extraction, 4) Data processing, 5) Event log generation, 6) Discovery of process models and 7) Evaluation of process models. The first two were already covered in Section 2 (MOOC definition) and at the beginning of section 5 (research questions). Steps 3 to 6 are described also in Section 5 and Section 6.

Within the literature, an extensive list of studies has addressed or comprehended the aforementioned challenges of self-directed learning. Notably, artificial intelligence is a highly favoured approach in this field. Specifically, classical data mining techniques, encompassing both supervised (Monllao Olive, Huynh, Reynolds, Dougiamas, & Wiese, 2020; Singh, Kumar, Bhushan, Kumar, & Vashishtha, 2021; Tang & Bao, 2022) and unsupervised (Ding, Yang, Yeung, & Pong, 2019; Ezen-Can, Boyer, Kellogg, & Booth, 2015; Soukaina, El Miloud, & Charaf, 2020) learning methodologies, find application in these research. However, despite such methods' prominence, process mining techniques remain conspicuously underutilised within this domain. This is intriguing, given the multiple advantages these techniques offer when navigating process-generated logs, a scenario frequently encountered within MOOC platforms.

Process mining has been used in a wide variety of fields of interest to researchers. For example, in industry, a comprehensive study identified the main challenges in business process management that can be addressed through process mining or other techniques (Beerepoot et al., 2023). Another study aims to optimise the procurement process from various perspectives using process mining (Van Der Aalst et al., 2007). In medicine, process mining enhances hospital information systems and their internal processes (Rojas, Munoz-Gama, Sepúlveda, & Capurro, 2016), and the appropriate treatment of clinical data through the application of process mining has improved the prediction of critical health outcomes in patients (Ashrafi, Abdollahi, Placencia, & Pishgar, 2024). In process optimisation, process mining has yielded highly accurate predictions of the estimated completion time for administrative procedures and bureaucratic processes (Van der Aalst, Schonenberg, & Song, 2011). Finally, in education (Bogarín et al., 2018), the patterns for the most and the least successful students are discovered using process mining techniques (Antonio Caballero-Hernández, Palomo-Duarte, Manuel Dodero, & Gaševic, 2024).

However, few studies have used process mining to address at least one of the challenges of self-directed learning. Among them, we can find works like (Maldonado-Mahauad et al., 2018; Matcha et al., 2019; Wong, Baars, de Koning, & Paas, 2021; Wong, Khalil, Baars, de Koning, & Paas, 2019) that use different process mining approaches to analyse the aforementioned self-regulation of learning.

Regarding the dropout and participation challenges, the literature covers these two challenges but to a lesser extent. For instance, (Deeva, Smedt, Koninck, & Weerdt, 2017; Rizvi, Rienties, & Rogaten, 2018) carry out studies that apply process mining to predict and understand why a user of a MOOC abandons their learning process. Other studies, like the ones developed by (Rizvi, Rienties, Rogaten, & Kizilcec, 2020; Okoye, Nganji, & Hosseini, 2019), focus their efforts on applying process mining techniques to discover variations in

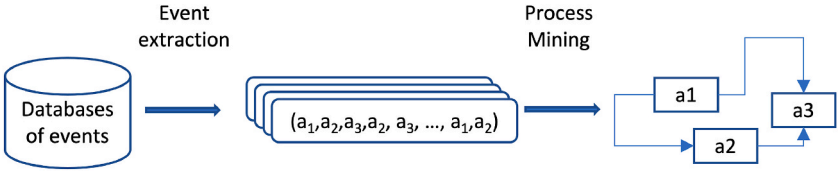


Fig. 1. Flow of an analysis using process mining.

the learning process user behaviour within a MOOC.

Actually, in the scarce literature that contemplates the challenges of self-directed learning, two related challenges have not aroused the interest of researchers: attention or interest and procrastination. Since process mining provides a visual insight into behavioural patterns within a large dataset of diverse users, this technique is particularly well suited to addressing the above challenges. Skilful data handling can identify these factors in the context of a well-defined process, such as a MOOC. In this study, an exploratory analysis of a group of users of a MOOC about language learning is carried out to find and understand patterns of procrastination or lack of interest in different groups of users. To the authors' knowledge, this is the first time procrastination patterns have been analysed using process mining.

4. MOOC data

The event log of the MOOC "Basic Spanish 1" has been used for this study. This course[1] is an introductory Spanish language MOOC offered by the edX platform and organised by the *Universitat Politècnica de València*. edX is an American MOOC provider created by Harvard and MIT in 2012. The MOOC is a 7-week Spanish language course aimed at users who want to learn conversational Spanish, starting with Spanish basics. The MOOC is free except for the certified diploma. This course is very popular since more than 500.000 users have finished this MOOC. The data was duly anonymised to avoid any possible identification of the users.

The course comprises seven modules of learning material, plus one week of introduction and two weeks of evaluation. Each module is made up of a set of lessons where a specific small component of educational material within the course is called a lesson. Each one of these lessons can be from one of the following five categories: reading, audio, video, pronunciation activity or assessment activity. Within the course structure, lessons are grouped into components called vertical components. The course has 32 readings in HTML format, 58 videos, 29 audios, 36 pronunciation activities and 89 evaluation activities.

In the mentioned course, 136,375 users were registered. However, only 96,815 developed activity, and they are the only ones considered in this study. Fig. 2 shows the flow of users through the weeks. The data of the enrolled users from the first edition of the course in the 2020 season was used.

For this study, the process mining technique was used to analyse the study sessions and procrastination levels of the users enrolled in the above-mentioned course. Therefore, it is necessary to define these concepts (process mining, study session and procrastination) to understand the context of this research.

4.1. Data extraction and processing

The data was extracted from the edX platform of the Universitat Politècnica de València, with the collaboration of his area of information and communication systems. The data was processed to eliminate the non-relevant users, remove the unnecessary columns, and unify the demographic, enrollment and course completion data. The main objective of this data processing is to retain the columns that allow for identifying and mapping the behaviour of the course users. Besides, users were categorised into two groups: verified users and audit users. The former were users who paid for a course certificate, and the latter were those who did not pay. Verified users need a minimum score of eight points to pass the course and obtain the certificate. Based on the minimum score, verified users were classified into three groups: Passed (with a score greater than or equal to eight), Failed (with a score less than eight), and Withdrawn (no score). Besides, the audited users were classified into two groups: Finished (users who finished the course) and Withdrawn (users who did not finish the course).

4.2. Study session

A study session can be defined as a set of chronologically ordered interactions with the components of a course in a given period of time (Vitiello, Walk, Helic, Chang, & Guetl, 2018). Researchers have identified or devised different ways of defining a study session within MOOC platforms. The easiest way is when the platform generates a unique identifier for each user session, as in the study of Yu et al. (Yu, Wu, & Liu, 2019) on the OpenEDU platform. Other studies define inactivity thresholds to determine the end of a session and the beginning of another. These time intervals are selected at the discretion of the researchers. This is the case of the studies (Ren, Rangwala, & Johri, 2016; Vitiello et al., 2018), which identify thresholds of 1 h and 30 min, respectively.

When the platform does not identify study sessions in its event logs, it is most accurate to determine an inactivity threshold to define the sessions and explicitly specify the reason for that decision. This is the case of the study (de Barba et al., 2020), where the threshold was determined based on the duration of the longest audiovisual material (approximately 23 min, plus an average reflection period of 7 min on the material learned).

4.3. Procrastination

Procrastination is a behaviour in which users tend to avoid completing tasks until the last minute. That is, they suspend learning time until they have no choice. This behaviour is related to poor academic performance in a traditional class (Khan, Schommer-Aikins,

[1] https://www.edx.org/course/basic-spanish-1-getting-started.

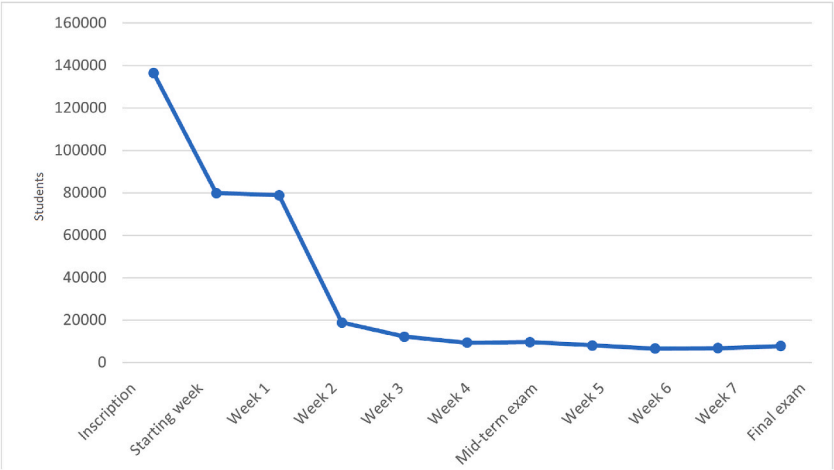


Fig. 2. Flow of users through the weeks.

& Saeed, 2021).

In the context of online learning environments, two forms of procrastination can be identified:

1. **Intra-session procrastination:** Characterised by intervals of inactivity during learning that are not wide enough to be considered as a threshold of inactivity between different study sessions.
2. **Extra-session procrastination:** Which is characterised by being an excessively wide interval of inactivity time between the end of one session and the beginning of another.

In this paper, the procrastination of the users of the MOOC of the case study will be analysed through the application of process mining techniques. An analysis of both intra-session procrastination and extra-session procrastination will be carried out. The methodology that has been used is detailed in the next section.

4.4. Event log generation

The generation of the event log aims to adapt the information to the possibility of answering research questions. With this in mind, this process clearly defines the actions of the users within a study session, with a clearly defined start and end, considering extended periods of inactivity within the study session and the time elapsed between study sessions of the same user. For this, the event log was generated in two stages: cleaning, where all unnecessary information was eliminated; and processing, where study sessions and behaviour of course users were identified.

For the processing stage, three types of language audiovisual material were identified: videos, audios and pronunciation activities.

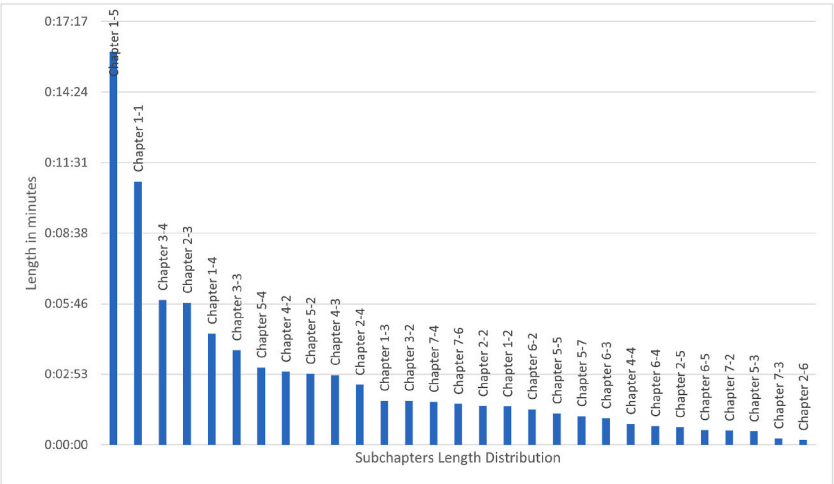


Fig. 3. Duration of audiovisual material by sub-chapter.

Generally, the longest audiovisual material is a video of 5 min and 5 s. However, each sub-chapter of the course uses the different types of audiovisual material identified. Consequently, all the durations of the audiovisual material need to be analysed as a whole, as Fig. 3 shows.

Based on the distribution of time in terms of sub-chapters (in Fig. 3 defined with the format chapter number - sub-chapter number), the one with the longest duration was 16 min with 3 s. Additionally, considering a reflection time such as that considered by (de Barba et al., 2020), the threshold of inactivity selected for this study was **25 min**.

Once the activity threshold was defined, the event log was labelled in study sessions for each user, each session being distant from its predecessor by periods of activity greater than or equal to 25 min. On the other hand, intra-session procrastination ranges from the duration of the longest individual audiovisual material (5 min and 5 s) to 24 min and 59 s.

Finally, time intervals from the two days of inactivity are considered for extra-session procrastination. Based on this classification, three types of extra-session procrastination were identified:

- Expected rest: from 25 min to 48 h of inactivity (for weekends).
- Ordinary procrastination: from 48 h to 10 weeks (expected course duration) of inactivity.
- Extraordinary procrastination: Rest intervals of more than 10 weeks duration.

5. Methodology

The research exposed in this paper tries to answer the following research questions:

1. What are the primary behavioural patterns exhibited by users within a study session in a MOOC concerning their interaction with the course components during study sessions?
2. How is procrastination related to the success of users in MOOCs?

The following subsections describe the process mining data analysis methodology developed (see Section 3).

5.1. Process models

This step applied process mining to discover process models used to address the research questions. The Disco Fluxicon software tool was used for this, which employs an adapted version of the Fuzzy Minner algorithm (Günther & Rozinat, 2012). In the following, we will describe how the event-logs have been processed to answer the research questions using process mining.

RQ1. What are the primary behavioural patterns exhibited by users within a study session in a MOOC concerning their interaction with the course components during study sessions?

The study sessions were identified in the event log to answer this question. These study sessions were defined at two hierarchical levels of the course structure:

1. The lesson components that represent each tab in a course sub-chapter.
2. The activity components (videos, problems, and drag-and-drop problems) that were parts of the lesson components.

Other activity components (like lectures, audios and pronunciations) are not labelled directly in the same way as videos, problems and drag-and-drop problems. In these cases, the interaction with the lesson components was considered as an interaction with its predominant activity component. For example, if the lesson component contains three audios and one short lecture, the lesson component was tagged like a section of audio. In this way, for this research question, two different event logs were used: one for lesson components and another one for activity components.

Summing up, the possible labels for lesson components would be: Pronunciation, Video, Audio, HTML, and Activity.

For logging events regarding lesson components, the features of interest are:

1. **modified:** Date and time the navigation log was updated.
2. **user_id_encrypt:** Unique encrypted identifier of the user.
3. **hierarchy:** Name of the hierarchical category of the component.
4. **vertical_component:** Predominant type of material in the lesson component.
5. **session:** Identifier of the user's study session in which the registered event took place.

To log events related to activity components, the features of interest are:

1. **module_type:** Type of study material.
2. **modified:** Date and time the navigation log was update.
3. **user_id_encrypt:** Unique encrypted identifier of the user.
4. **hierarchy:** Hierarchical position of the visited component in the course structure.
5. **session:** Identifier of the user's study session where the registered event occurred.

RQ2. How is procrastination related to the success of users in MOOCs?

To answer this question, the lesson components were used. Study sessions were identified for all users. In these, the pauses of 10 min or more were tagged with the code "distraction". The event logs used in this research question are made up of six columns, described below:

1. **modified:** Date and time the navigation log was updated.
2. **user_id_encrypt:** Unique encrypted identifier of the user.
3. **hierarchical:** Hierarchical position of the visited component in the course structure.
4. **type_content:** Type of study material.
5. **session:** Identifier of the user's study session where the registered event occurred.
6. **attention:** A two-state identifier that determines whether there was a distraction period between two events in the same study session.

6. Evaluation of process models

This section evaluates and interprets the process models obtained from the log events resulting from the last subsection. These evaluations are organized based on the research questions. For each research question, one or more process models were generated (see Section 3). The displayed process models are absolute frequency graphs, where the numerical information in each process model generally represents the most common interactions for each node and the most frequent transitions between nodes recorded in the analysed data. The complete process models obtained for this study are available in Appendix A.

The absolute frequency graphs provide an overview of the most common and less common interactions among different components of the analysed data. In fact, the graphs allow us to visually understand how users interact with different components of the course and gain a visible perspective on the impact of each component in the context of a MOOC. In this way, the resulting graphs represent users' behaviour across study sessions. In each scenario, the numbers present in the nodes and edges represent the number of times the users perform the activity (for nodes) and the number of times the users have transitioned from one activity to another (for edges). The numbers do not add up as infrequent activities and transitions have been removed to increase the clarity of the graph. In any case, the density or "popularity" of each activity or transition for each type of user is important for the pattern analysis.

In our initial analysis, we examined the activity components identified by the edX platform for both R1 and R2 results. In the subsequent phases, we will delve deeper into identifying the specific components of the learning process based on the course design.

6.1. RQ1. What are the primary behavioural patterns exhibited by users within a study session in a MOOC concerning their interaction with the course components?

The event log used for this research question resulted in the process models presented in Figs. 4–6. The process models delineated in the aforementioned figures encapsulate users' conduct within a study session, spanning the specified time intervals specified in section 4.4. The analysis carried out on these process models concluded in five results (R) detailed below:

6.1.1. R1. Study sessions of passed verified users provide a complementary path between the material and the assessments

Verified users, who have passed the course, typically complete a study session between at least two activity components (video,

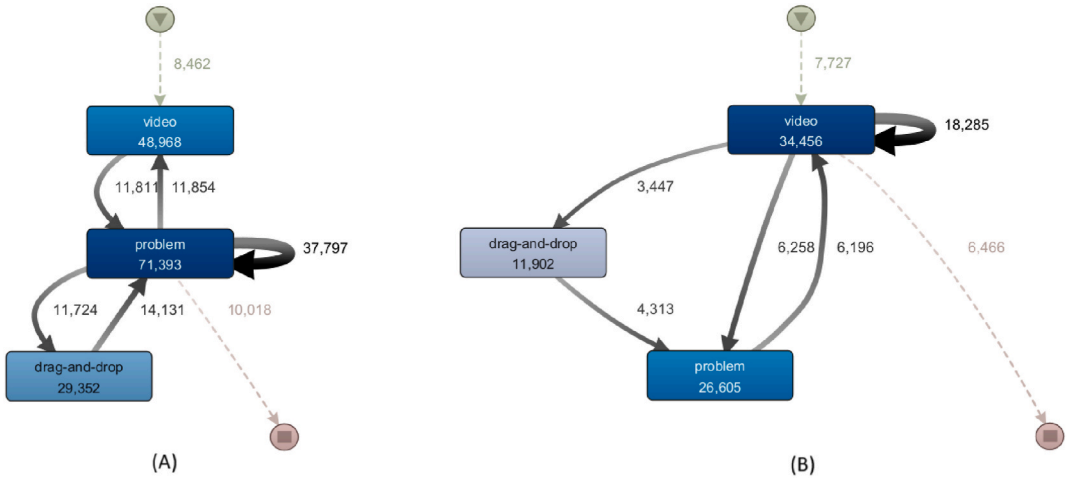


Fig. 4. Behaviour of verified users who passed (a) or failed (b) in terms of study sessions.

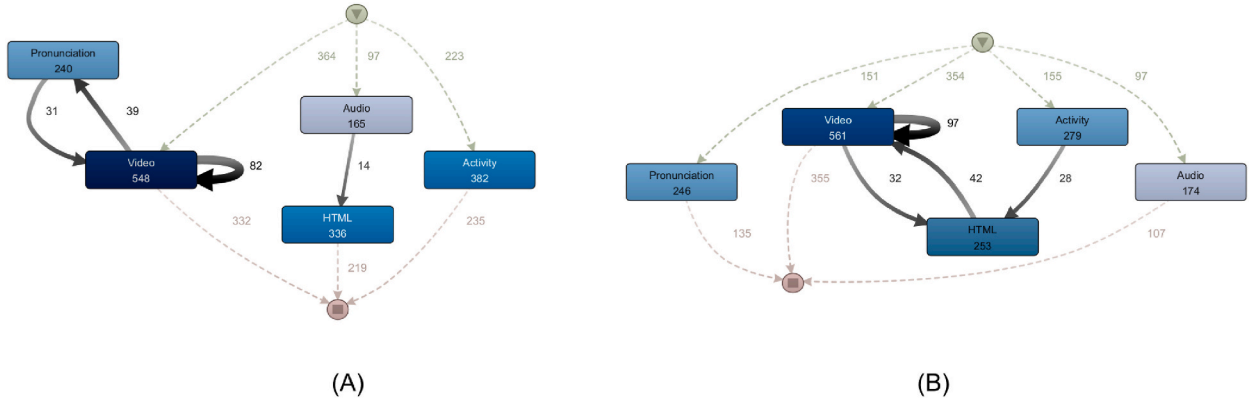


Fig. 5. Behaviour of verified users who passed (a) or failed (b) in the lesson components.

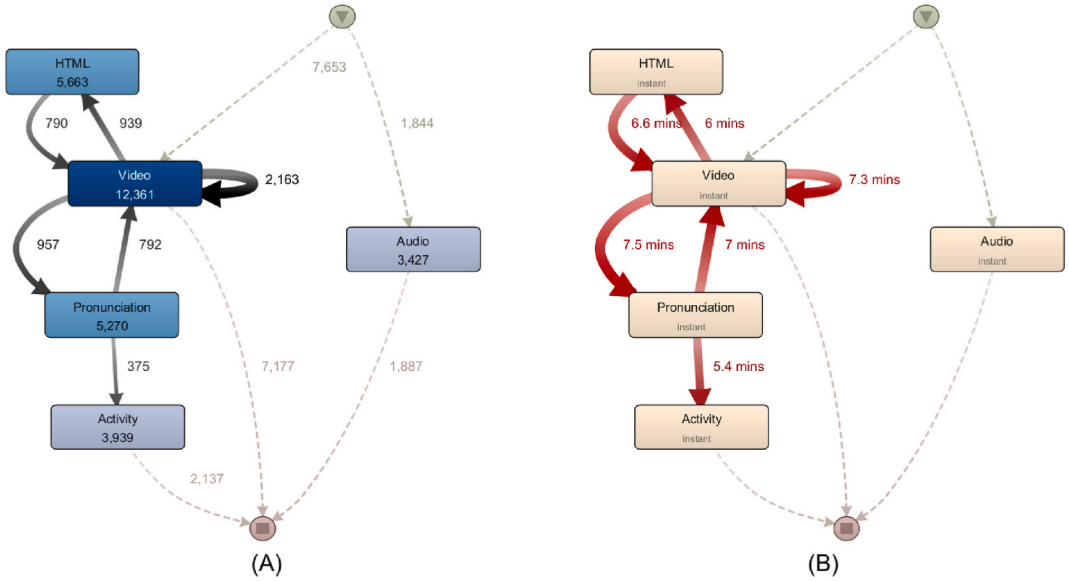


Fig. 6. Behaviour of audit users who finished the course based on lesson components. In (a), we show the number of transactions between nodes, while in (b), we include the travel time between nodes.

problem, or drag-and-drop problem). Fig. 4 shows the behaviour by study session of the passed and failed verified users. As can be seen, successful users start their study sessions with videos, but they also include different activities. On the other hand, failed users mostly use videos in their study sessions. Furthermore, failed users frequently work in sessions where they access only one video before concluding the session.

6.1.2. R2. Verified users who failed tend to look for quiz answers in the content before responding

Passed verified users have studied sessions that involve different types of activities in a loop, as can be seen in the left graph of Fig. 4. This behaviour shows a constant sequence in course evaluations by these passed verified users. However, failed verified users try to find the answer to the questions by searching the material when they try to take the quizzes (right graph).

6.1.3. R3. Passed verified users complete study sessions focused on solving the assessments

In the lesson components event log, passed verified users carry out study sessions only for evaluations. However, failed verified users do another type of behaviour: they do not interact only with the assessment components, but try to find the results of the questions in the course material (videos, lectures) during these assessments (see Fig. 5). This complements result two (R2), where the process model shows that failed verified users try to find the answer to the quizzes while they are solving them.

6.1.4. R4. Verified users make better use of listening and pronunciation material

As illustrated in Fig. 5, verified users who have successfully passed demonstrate more proficient use of audio and pronunciation

materials. The findings indicate that these successful users engage in study sessions where supplementary materials are incorporated alongside the pronunciation exercises. On the contrary, verified users who did not succeed tend to concentrate solely on pronunciation exercises during separate study sessions.

6.1.5. R5. Audit users conduct reading and audiovisual material review sessions before accessing the available activities, but tend to conduct separate audio sessions

Audit users exclude audio lessons from their study sessions. In Fig. 6, we can see how the study sessions of the audit users considered practically all the different types of lessons, only the audio lessons were reviewed in separate sessions, and the duration of the audio study sessions was short. Based on this, it can be induced that audit users are not attracted by audio lessons.

6.2. RQ2. How is procrastination related to the success of users in MOOCs?

In conducting this analysis, certain activities within the sequence of events have been subject to relabelling (see Fig. 1). Specifically, activities involving procrastination, which exceed the intra-session time threshold, have been assigned the "Distraction" suffix. This adjusted set of relabelled event sequences is the foundation for constructing the process models depicted in Fig. 7. The research conducted on these process models has yielded four outcomes (R) to be elaborated on in the ensuing discussion:

6.2.1. R6. The pattern of intra-session procrastination is similar in passed and failed users

Fig. 7 (*a* and *b*) shows the behaviour patterns of verified users (both passed and failed). The video sequence is able to maintain attention in a study session. In both cases, procrastination occurs when a change in a study lesson is required. The major difference between the two groups is found in the pattern of activities per se, and not in the presence or absence of procrastination between activities. This is because all user groups procrastinate on some activity at some point. Regardless of the time interval between activity sequences, what truly influences the results obtained in this study is the sequence of activities exhibited by different users before and

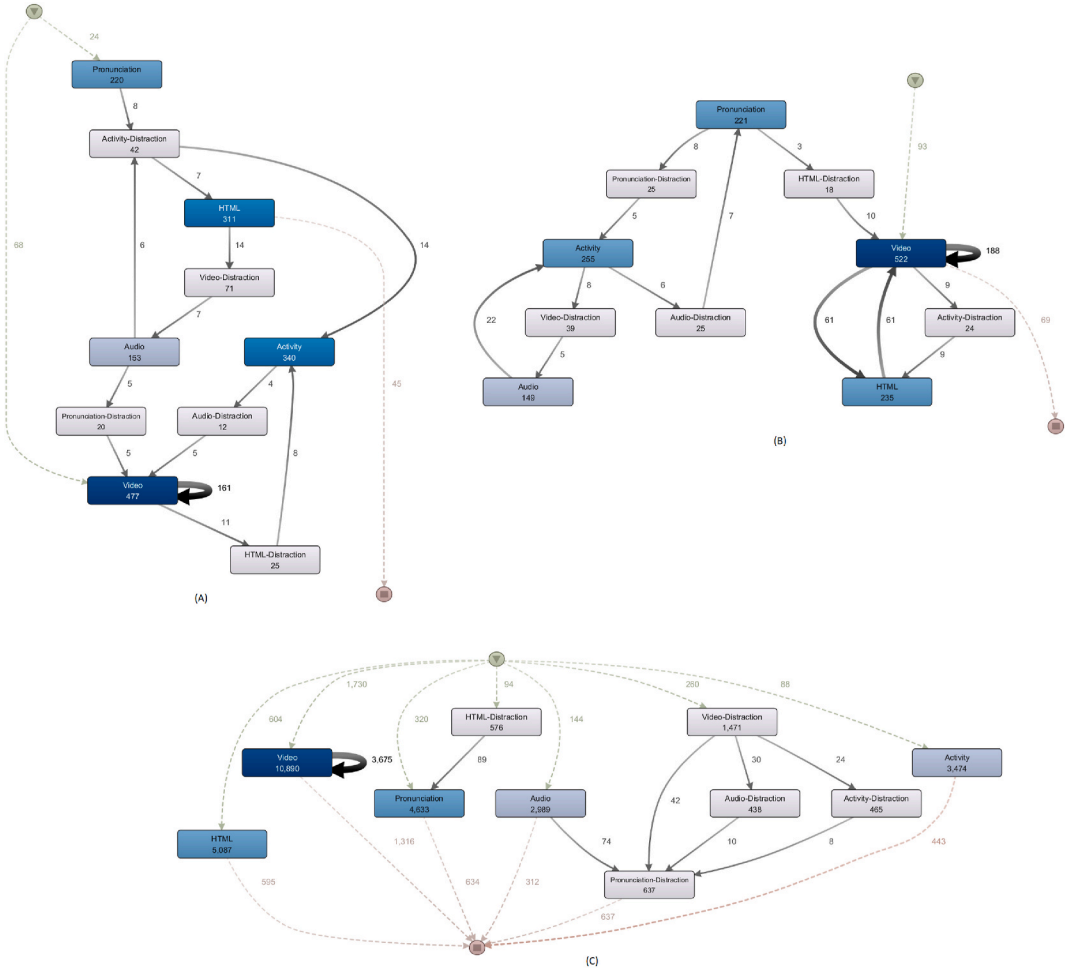


Fig. 7. Attention of verified users who passed (a) and failed (b). Audit users who complete the course (c).

after procrastination.

6.2.2. R7. The presence of intra-session procrastination is much more pronounced in the case of audit users

As seen in Fig. 7 c, intra-procrastination is more common in the case of audit users. As can be seen, there are complete sessions in which this type of user procrastinates in all activities. In fact, audit users procrastinate whenever the study session involves more than one activity.

6.2.3. R8. Extra-session procrastination could have some influence on the success or failure of users in the course

For extra-session procrastination, all event records at all levels were considered to analyse all possible interactions before calculating the time between sessions. Based on this, three distribution functions were analysed:

- The expected rest distribution function for verified users.
- The ordinary procrastination distribution function for verified users.
- The extraordinary Procrastination distribution function for verified users

Distribution graphs, also known as probability distribution graphs or frequency distribution graphs, are visual representations of the distribution of data values in a dataset. The area under the curve is always one, and its intervals can be interpreted as the percentage of users in that rest interval.

Fig. 8 shows the expected rest distribution function for verified users. As can be seen, the behaviour of passed verified and failed verified users in scheduled break intervals is similar. The curve tends to be steeper to the left in all three cases. However, for users who dropped out of the course, the curve highest peak is wider, which evidences more varied time intervals to finish one session and start another.

With respect to ordinary procrastination, Fig. 9 shows the corresponding distribution function for all three groups of verified users. As can be seen, they have a curve that tends to be steeper to the left. These curves do not have as sharp peaks as the curves in the expected rest. However, the curves of the failed and withdrawn users are higher to the right. This means that the duration of ordinary procrastination of the failed and withdrawn users is higher than that of the passed users.

Finally, Fig. 10 shows the extraordinary procrastination distribution function for verified users. As we have previously defined, extraordinary procrastination reflects periods of inactivity between the end of a session and the beginning of another greater than the duration of the course. As Fig. 10 shows, the three groups of users present this behaviour in periods of time that are close to 500 h between sessions since the peak of their curves is to the left. However, the curve of withdrawn users tends to be higher as it goes to the right.

7. Limitations

The findings of this study need to be considered in the context that there are limitations to relying solely on event log activity to understand the experiences of MOOC participants. While event logs provide valuable insights into the system's perspective and can offer useful data for platform improvement, they do not capture the nuanced and multifaceted nature of participants' learning journeys. The categorisation of learners as 'passed', 'failed' or 'dropped' is limited to reflect the different goals, motivations and behaviours of MOOC participants.

One key aspect to consider is the voluntary nature of MOOC enrollment. Unlike traditional educational settings where users may be compelled to complete courses to fulfil academic requirements, MOOC participants often join out of personal interest or self-driven learning goals. Consequently, the strict categorisation based on course completion may not represent the intentions and expectations of these learners.

Furthermore, external factors and competing priorities can influence participants' ability to continue with a MOOC. Many learners have professional or personal commitments that may restrict their time and availability. Thus, they might discontinue or temporarily pause their participation, leading to a 'dropped' classification. These individuals should not be regarded as failures or disengaged learners but rather as individuals who faced constraints and had to prioritize other responsibilities. Acknowledging the influence of

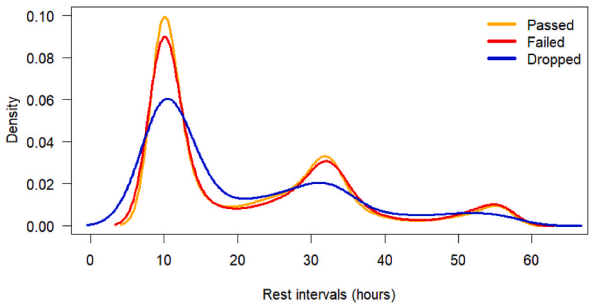


Fig. 8. Expected rest distribution function for verified users.

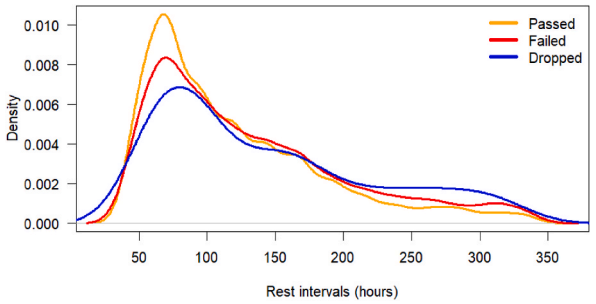


Fig. 9. Ordinary procrastination distribution function for verified users.

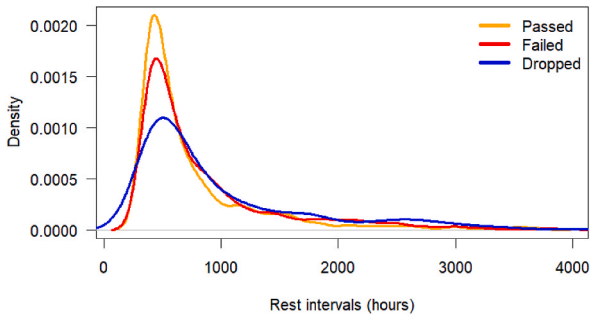


Fig. 10. Extraordinary procrastination distribution function for certified users.

external factors is essential to avoid misinterpretations of the event log data and to understand the participants' experiences better.

Additionally, the dichotomy between those seeking certification and those enrolling for personal enrichment introduces another layer of complexity. Learners who pay for certification may have different motivations and engagement levels compared to those who do not seek formal recognition (Littenberg-Tobias, Ruipérez-Valiente, & Reich, 2020). The former group might be more motivated to complete the course and adhere to the prescribed structure, as they have invested financial resources and expect tangible outcomes. In contrast, learners focused on personal enrichment may be more flexible in their learning path, selecting specific modules or skipping sections they find redundant due to prior knowledge. Such variations in behaviour should be considered while interpreting the event log data and understanding the reported experiences of participants.

A hybrid approach could be employed to bridge the gap between event logs and participant experiences. By combining quantitative data from event logs with qualitative data derived from participant surveys, interviews, or open-ended feedback, researchers can better understand the factors influencing learners' behaviours and subjective experiences. This holistic approach could enable the researchers to uncover patterns, motivations, and learning trajectories that may not be apparent solely from event log analysis.

Therefore, relying solely on event logs to understand MOOC participants' experiences has limitations in terms of categorisation and the diversity of learner goals and behaviours (Bogarín et al., 2018). Despite these limitations, research conducted using event logs can still improve MOOC users' performance.

In this sense, procrastination presents a multifaceted challenge that surpasses the realm of user engagement solely within the course material. Numerous other user-related factors can trigger this behaviour. Identifying procrastination within the course tracking process is the preliminary phase of addressing this challenge. The pivotal aspect entails formulating effective techniques that facilitate continuous user engagement with the course and its content, considering that this engagement remains unaffected by fixed schedules or predefined compliance conditions.

Finally, it should be noted that event log analysis also helps identify areas for improvement in course structure and content delivery, which is the main goal of the research described in this paper. Thus, while acknowledging the limitations of this research, we strongly believe that using event log data can improve MOOC users' learning experience and outcomes.

8. Conclusions

MOOC platforms are technological tools that have transformed learning. These tools allow their users to access courses from anywhere and anytime. However, their low completion rates have caught the attention of researchers worldwide. Several researchers have focused their efforts on analysing the different challenges faced by users of MOOCs. There are a large number of studies that have focused on dropout, participation and self-regulation; while other challenges such as attention and procrastination have received little attention.

Process mining is a discipline of artificial intelligence that analyses data generated from interactions. This discipline has a set of

advantages, among which are: 1) real-time modelling and analysis of events, 2) modelling processes without the need of a base model (i.e., the model is generated only from the data), 3) easily comparing a set of process models in similar circumstances and 4) clear visualization of the behaviour to be analysed.

This study conducted an analysis focused on finding procrastination patterns in a MOOC of language learning by applying process mining discovery techniques. Among the various conclusions obtained, focused on the process mining graphs constructed, the following stand out:

1. Users who stay and finish the course have utilized the MOOC material better as the process graphs illustrate. Thus, interest in learning from the MOOC material is a contributing factor for continuing and completing the course.
2. In relation to the above, if we compare the behaviour of users who pass the course with those who do not, we can see that the latter make inadequate use of the supplementary materials. Thus, it can be observed that users who do not pass the course tend to combine evaluation tasks with consultation of the support materials, which seems to indicate that they are trying to identify and respond correctly to the activities without having absorbed the essential knowledge.
3. In relation to the time intervals devoted to each of the study sessions, as graphs show, users who pass the course show a greater ability to optimally engage with the varied contents of the activity components (Figs. 4 and 5). This is evident in the pattern of behaviour extracted using PM. Thus, it is observed that users who pass the course develop a coherent sequence between activities including the evaluation phases. This behavioural observation highlights a clear relationship between the appropriate use of the course contents and the subsequent acquisition of knowledge, which is ultimately reflected in the results.
4. We have also observed that intra-session procrastination patterns (Fig. 7) are closely linked to users' course outcomes, whether they pass or fail. The results show a notable disparity: users who successfully pass the course show fewer instances of intra-session procrastination than their unsuccessful counterparts. Successful users quickly redirect their attention to the videos when their concentration wanes. Conversely, unsuccessful users tend to navigate through the material without revisiting the content, indicating a distinct pattern of engagement.

Other conclusions that can be extracted from the study are:

1. The considerable drop in course participation during the second week of the MOOC underscores the imperative of implementing effective strategies to retain users and counter this trend. In the context of a language-learning MOOC encompassing a diverse array of activity components, it becomes evident that content diversity, while valuable, is insufficient to ensure user engagement throughout the course. Therefore, it is essential to devise content structures that are more engaging and compelling, capturing users' interest and commitment to successfully completing the course.
2. Finally, we can see that extra-session procrastination could impact course results. The analysis of the distribution of course users' extra-session procrastination (Figs. 9 and 10) shows that there are differences that are important in the case of the verified users to succeed in the course, especially when we look at ordinary extra-session procrastination.

In general terms, the results obtained in this study show that procrastination is related to the success or failure of the course, but it is not the only relevant factor. The type of material analysed and how the content is explored during the study sessions are also related to the final result.

This study is somewhat limited in reflecting the different goals, motivations and behaviours of MOOC participants. The division may be somewhat artificial based on the data alone. Therefore, it would be interesting to include two surveys in the courses to be conducted. A previous one, in which the user of the MOOC can mark his/her objectives and a final one in which the user can indicate the reasons for his/her success, failure and/or abandonment of the course. With all this information, it would be possible to better classify and analyse user behaviour.

In summary, this paper presents a method based on process mining and procrastination analysis that can detect the possible reasons why a user fails a MOOC. As we have shown, procrastination and behaviour during study sessions directly influence users' final performance. Although our study has been limited to one specific MOOC, we believe the methodology can be easily extrapolated to other MOOCs using the same platform (edX) since the process mining method used in this work is basically based on analysing the learning platform logs. The deployment of analytical tools, such as the one we propose in this work, in learning platforms such as edX could allow instructors of online courses to improve the tracking of learners. In this way, they could easily detect learners who exhibit behaviours that may lead them to fail or drop out of the course and, in such situations, propose actions to recover those learners.

Funding

We thank the anonymous reviewers for their comments. This work has been funded by ValGrai, CIPROM/2022/6 (FASSLOW) and IDIFEDER/2021/05 (CLUSTERIA) funded by Generalitat Valenciana and PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "FEDER: a way to make Europe". Finally, we would like to thank the Department of Information and Communication Systems (ASIC) of the Polytechnic University of Valencia, with special mention to Ignacio Despujol, for providing us with access to the information and data used during the preparation of this article.

CRediT authorship contribution statement

Jorge A. Vázquez Mendoza: Writing – original draft, Methodology, Formal analysis, Data curation. **Cèsar Ferri Ramírez:** Writing – review & editing, Writing – original draft, Supervision. **Carlos Monserrat Aranda:** Writing – review & editing, Writing – original draft, Supervision.

Data availability

Data will be made available on request.

A Appendix A. Original process models

The compressed file with the original process models can be found here (shorturl.at/lqOX8).

References

Aljaraideh, Y. (2019). Massive open online learning (MOOC) benefits and challenges: A case study in Jordanian context. *International Journal of Instruction, 12*, 65–78.

Antonio Caballero-Hernández, J., Palomo-Duarte, M., Manuel Dodero, J., & Gaševic, D. (2024). Supporting skill assessment in learning experiences based on serious games through process mining techniques. *International Journal of Interactive Multimedia & Artificial Intelligence, 8*.

Ashrafi, N., Abdollahi, A., Placencia, G., & Pishgar, M. (2024). Effect of a process mining based pre-processing step in prediction of the critical health outcomes. *arXiv preprint, 1*, 1–2. arXiv:2407.02821.

Beerepoot, I., Di Ciccio, C., Reijers, H. A., Rinderle-Ma, S., Bandara, W., Burattin, A., et al. (2023). The biggest business process management problems to solve before we die. *Computers in Industry, 146*, Article 103837.

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*, Article e1230.

Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2019). Predict and intervene: Addressing the dropout problem in a MOOC-based program. In *Proceedings of the sixth (2019) ACM conference on learning@ scale* (pp. 1–9).

Cagiltay, N. E., Toker, S., & Cagiltay, K. (2023). Exploring the influence of countries' economic conditions on massive open online course (MOOC) participation: A study of 3.5 million mitx learners. *International Review of Research in Open and Distributed Learning, 24*, 1–17.

de Barba, P. G., Malekian, D., Oliveira, E. A., Bailey, J., Ryan, T., & Kennedy, G. (2020). The importance and meaning of session behaviour in a MOOC. *Computers & Education, 146*, Article 103772.

Deeva, G., Smedt, J. D., Koninck, P. D., & Weerdt, J. D. (2017). Dropout prediction in MOOCs: A comparison between process and sequence mining. In *International conference on business process management* (pp. 243–255). Springer.

Ding, M., Yang, K., Yeung, D. Y., & Pong, T. C. (2019). Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 135–144).

Diver, P., & Martinez, I. (2015). MOOCs as a massive research laboratory: Opportunities and challenges. *Distance Education, 36*, 5–25.

Doss, A. N., Krishnan, R., Karuppasamy, A. D., & Sam, B. (2024). Learning analytics model for predictive analysis of learners behavior for an indigenous MOOC platform (tadakhul system) in Oman. *International Journal of Information and Education Technology, 14*.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146–150).

Günther, C. W., & Rozinat, A. (2012). Disco: Discover your processes. *BPM (Demos), 940*, 40–44.

Huang, N., Zhang, J., Burtch, G., Li, X., & Chen, P. (2021). Combating procrastination on massive online open courses via optimal calls to action. *Information Systems Research, 32*, 301–317.

Jarnac de Freitas, M., & Mira da Silva, M. (2023). Systematic literature review about gamification in MOOCs. *Open Learning: The Journal of Open, Distance and e-Learning, 38*, 73–95.

Khan, I., Schommer-Aikins, M., & Saeed, N. (2021). Cognitive flexibility, procrastination, and need for closure predict online self-directed learning among pakistani virtual university students. *International Journal of Distance Education and E-Learning, 6*, 31–41.

Littenberg-Tobias, J., Ruipérez-Valiente, J. A., & Reich, J. (2020). Studying learner behavior in online courses with free-certificate coupons: Results from two case studies. *International Review of Research in Open and Distributed Learning, 21*, 1–22.

Liu, Z., Tang, Q., Ouyang, F., Long, T., & Liu, S. (2024). Profiling students' learning engagement in MOOC discussions to identify learning achievement: An automated configurational approach. *Computers & Education, 219*, Article 105109.

Maldonado, J. J., Palta, R., Vázquez, J., Bermeo, J. L., Pérez-Sanagustín, M., & Munoz-Gama, J. (2016). Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach. In *2016 XLII Latin American computing conference (CLEI)* (pp. 1–12). IEEE.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Delgado-Kloos, C. (2018). Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning. In *European conference on technology enhanced learning* (pp. 355–369). Springer.

Matcha, W., Gašević, D., Uzir, A., Jovanović, J., Pardo, A., Maldonado-Mahauad, J., et al. (2019). Detection of learning strategies: A comparison of process, sequence and network analytic approaches. In *European conference on technology enhanced learning* (pp. 525–540). Springer.

Mesquita, M. A., Toda, A. M., & Brancher, J. D. (2014). Brasileduca—an open-source MOOC platform for Portuguese speakers with gamification concepts. In *2014 IEEE frontiers in education conference (FIE) proceedings* (pp. 1–7). IEEE.

Monllao Olive, D., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2020). A supervised learning framework: Using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education, 32*, 9–26.

Mrhar, K., Douimi, O., & Abik, M. (2021). A dropout predictor system in MOOCs based on neural networks. *Journal of Automation, Mobile Robotics and Intelligent Systems*, 72–80.

Mutawa, A. M. (2023). *Perspective chapter: MOOCs at higher education: Current state and future trends*.

Nieves, D., Ramírez-Quintana, M. J., Monserrat, C., Ferri, C., & Hernández-Orallo, J. (2020). Learning alternative ways of performing a task. *Expert Systems with Applications, 148*, Article 113263.

Norman, H., & Nordin, N. (2024). Construction of a sentiment analysis model for Chinese MOOC comments based on big data. *Educational Administration: Theory and Practice, 30*, 10186–10190.

Okoye, K., Nganji, J. T., & Hosseini, S. (2019). Learning analytics: The role of information technology for educational process innovation. In *International conference on innovations in bio-inspired computing and applications* (pp. 272–284). Springer.

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Exploring the use of MOOC discussion forums. In *Proceedings of London international conference on education, LICE* (pp. 1–4).

Read, T. (2014). 6 the architectonics of language MOOCs. In *Language MOOCs* (pp. 91–105). De Gruyter Open Poland.
Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting performance on MOOC assessments using multi-regression models. *arXiv preprint arXiv:1605.02269*.
Rizvi, S., Rienties, B., & Rogaten, J. (2018). Temporal dynamics of MOOC learning trajectories. In *Proceedings of the first international conference on data science, E-learning and information systems* (pp. 1–6).
Rizvi, S., Rienties, B., Rogaten, J., & Kizilcec, R. F. (2020). Investigating variation in learning processes in a futurelearn MOOC. *Journal of Computing in Higher Education, 32*, 162–181.
Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics, 61*, 224–236.
Shah, D. (2021). MOOC stats 2021. URL: https://www.classcentral.com/report/mooc-stats-2021/.
Singh, A. K., Kumar, S., Bhushan, S., Kumar, P., & Vashishtha, A. (2021). A proportional sentiment analysis of MOOCs course reviews using supervised learning algorithms. *Ingénierie des Systèmes d'Information, 26*.
Song, L., & Hill, J. R. (2007). A conceptual model for understanding self-directed learning in online environments. *The Journal of Interactive Online Learning, 6*, 27–42.
Soukaina, S., El Miloud, S., & Charaf, M. E. H. (2020). MOOCs performance analysis based on quality and machine learning approaches. In *2020 IEEE 2nd international conference on electronics, control, optimization and computer science (ICECOCS)* (pp. 1–5). IEEE.
Tang, H., & Bao, Y. (2022). Self-regulated learner profiles in MOOCs: A cluster analysis based on the item response theory. *Interactive Learning Environments*, 1–17.
Teixeira, A. M., & Mota, J. (2014). A proposal for the methodological design of collaborative language MOOCs. *Language MOOCs: Providing learning, transcending boundaries*, 33–47.
Thiede, M., Fuerstenau, D., & Barquet, A. P. B. (2018). How is process mining technology used by organizations? A systematic literature review of empirical studies. *Business Process Management Journal*.
Van Der Aalst, W. M., Reijers, H. A., Weijters, A. J., van Dongen, B. F., De Medeiros, A. A., Song, M., et al. (2007). Business process mining: An industrial application. *Information Systems, 32*, 713–732.
Van der Aalst, W. M., Schonenberg, M. H., & Song, M. (2011). Time prediction based on process mining. *Information systems, 36*, 450–475.
Van Eck, M. L., Lu, X., Leemans, S. J., & Van Der Aalst, W. M. (2015). PM2: A process mining project methodology. In *International conference on advanced information systems engineering* (pp. 297–313). Springer.
Vázquez Mendoza, J. A. (2021). *Análisis y descubrimiento de patrones de comportamiento en estudiantes de cursos online*. Universitat Politècnica de València.
Vitiello, M., Walk, S., Helic, D., Chang, V., & Guetl, C. (2018). User behavioral patterns and early dropouts detection: Improved users profiling through analysis of successive offering of MOOC. *Journal of Universal Computer Science, 24*, 1131–1150.
Wang, W., Zhao, Y., Wu, Y. J., & Goh, M. (2022). Factors of dropout from MOOCs: A bibliometric review. *Library Hi Tech*.
Wong, J., Baars, M., de Koning, B. B., & Paas, F. (2021). Examining the use of prompts to facilitate self-regulated learning in massive open online courses. *Computers in Human Behavior, 115*, Article 106596.
Wong, J., Khalil, M., Baars, M., de Koning, B. B., & Paas, F. (2019). Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. *Computers & Education, 140*, Article 103595.
Yao, M., Sahebi, S., & Feyzi Behnagh, R. (2020). Analyzing student procrastination in MOOCs: A multivariate hawkes approach. In *Proceedings of the 13th conference on educational data mining (EDM2020)*.
Yu, C. H., Wu, J., & Liu, A. C. (2019). Predicting learning outcomes with MOOC clickstreams. *Education Sciences, 9*, 104.