Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs

Pedro M. Moreno-Marcos, Pedro J. Muñoz-Merino, Jorge Maldonado-Mahauad, Mar Pérez-Sanagustin, Carlos Alario-Hoyos, Carlos Delgado Kloos, José A Ruipérez-Valiente, Thomas Staubitz, Matt Jenner, Sherif Halawa, et al.

Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs

ABSTRACT

MOOCs (Massive Open Online Courses) have usually high dropout rates. Many articles have proposed predictive models in order to early detect learners at risk to alleviate this issue. Nevertheless, existing models do not consider complex high-level variables, such as self-regulated learning (SRL) strategies, which can have an important effect on learners' success. In addition, predictions are often carried out in instructor-paced MOOCs, where contents are released gradually, but not in self-paced MOOCs, where all materials are available from the beginning and users can enroll at any time. For self-paced MOOCs, existing predictive models are limited in the way they deal with the flexibility offered by the course start date, which is learner dependent. Therefore, they need to be adapted so as to predict with little information short after each learner starts engaging with the MOOC. To solve these issues, this paper contributes with the study of how SRL strategies could be included in predictive models for self-paced MOOCs. Particularly, self-reported and event-based SRL strategies are evaluated and compared to measure their effect for dropout prediction. Also, the paper contributes with a new methodology to analyze self-paced MOOCs when carrying out a temporal analysis to discover how early prediction models can serve to detect learners at risk. Results of this article show that event-based SRL strategies show a very high predictive power, although variables related to learners' interactions with exercises are still the best predictors. That is, event-based SRL strategies can be useful to predict if e.g., variables related to learners' interactions with exercises are not available. Furthermore, results show that this methodology serves to achieve early powerful predictions from about 25-33% of the theoretical course duration. The proposed methodology presents a new approach to predict dropouts in self-paced MOOCs, considering complex variables that go beyond the classic trace-data directly captured by the MOOC platforms.

Keywords

Data science applications in education, Distance education and online learning, Lifelong learning, Post-secondary education

1. INTRODUCTION

Due to their openness nature, MOOCs (Massive Open Online Courses) have attracted a wide diversity of learners. In fact, learners have seen MOOCs as an opportunity to learn from anywhere at their own pace. However, most of the learners who enroll in MOOCs do not manage to finish them. Completion rates in MOOCs typically fall below 10% (Daniel, 2012). As a consequence, there have been lot of research on predictive models in the last years to forecast learning outcomes, such as dropout and scores, (e.g., Feng, Tang, & Liu, 2019) and learners' behaviors (Bote-Lorenzo & Gómez-Sánchez, 2017).

Predictive models provide timely information about learners at risk of dropout in order to inform interventions. Both, teachers and learners, can benefit from the results of the predictive models (blinded). For example, teachers can adapt their courses to engage and support their learners, by providing them with personalized support; while learners can get insights and receive feedback of their learning process. Particularly, predictive models can serve to generate alerts for instructors and learners, adapt the contents, motivate their learners with information about how they are doing, and improve the curriculum design, among others (Cui, Chen, Shiri, & Fan, 2019).

Despite these benefits, current predictive models typically face some limitations. First, predictive models are often carried out at the end of the MOOC as a post-hoc analysis, so they fail to anticipate learners at risk. Because of that, more studies are needed to identify when is the best moment to predict to achieve early prediction. Nevertheless, temporal thresholds about the best moment to predict are context-dependent and there is always a trade-off between anticipation and predictive power (blinded) that should be further explored. Furthermore, although these timing considerations have already been considered for instructor-paced MOOCs (e.g., Ruipérez-Valiente, Cobos, Muñoz-Merino, Andújar, & Delgado Kloos, 2017; Xing, Chen, Stein, & Marcinkowski, 2016), they have not been addressed in self-paced MOOCs, where learners have the flexibility to follow the course at their own pace because all materials are released at the

beginning. That is, learners are not synchronized in the course and thus, it is not possible to use predictive models in particular periods of time (i.e., at a specific date, each learner can be engaging with different parts of the course, and even a learner may have not started yet or have already finished the course).

Another limitation of current predictive models is that they are usually developed taking as predictors simple events, such as course interactions with e.g., videos (Ye, et al., 2015) and exercises (Boyer & Veeramachaneni, 2015), and click-stream activity (Halawa, Greene, & Mitchell, 2014). While there have been accurate results of predictive models using low-level events (e.g., Ruipérez-Valiente et al., 2017; Laveti, Juppili, Ch, Pal, & Babu, 2017), these models may behave better if they include more complex behaviors such as sequences of learners' behavior patterns (i.e., learning strategies) when interacting with course contents (hereafter called sequence patterns). In self-paced MOOCs, learners are less guided so there is a wider variability of interactions with the platforms that lead to different interactive patterns. Given this variability, predictive models could benefit from including these sequence patterns as a new variable. New studies are required to analyze if these patterns add value to the predictive models in self-paced MOOCs.

Research has shown that completion rates are usually lower in self-pace settings (not only MOOCs) (Rhode, 2009). Lack of self-regulated learning (SRL) skills can make it difficult to succeed in this unguided learning environments (blinded). Self-regulation is the process by which learners take control of their learning process to achieve the proposed learning goals (blinded). Prior works have shown that the lack of SRL skills can be an important factor that leads to failure and dropout in MOOCs (Terras & Ramsay, 2015), and particularly in self-paced MOOCs (blinded). In this line, research has showed that SRL strategies such as time management, motivation, self-monitoring, etc., have impact on learners' success (Kizilcec & Schneider, 2015; Zheng, Rosson, Shih, & Carroll, 2015). However, previous articles often use only self-reported SRL strategies (e.g., blinded; Hood, Littlejohn, & Milligan, 2015) and not the actual strategies obtained from learners' interactions (event-based strategies, such as SRL sequence patterns, which we initially explored in a previous contribution, (blinded), and they do not evaluate how SRL strategies can have an impact as predictors for dropout.

In this context, the aim of this work is to analyze whether SRL strategies can improve current dropout predictive models, with or without other common self-reported variables and variables obtained from click-stream data. With this aim, and considering the importance of predicting on time, the following research questions are addressed in this paper:

RQ1: What is the predictive power of self-reported SRL strategies in dropout prediction?

RQ2: What is the predictive power of event-based SRL strategies in dropout prediction?

RQ3: When is the best moment to predict dropout in a self-paced MOOC?

In order to address these questions, we follow an analytical method based on data from a self-paced MOOC, with a validation through the analysis of other self-paced MOOCs.

2. RELATED WORK

2.1. PREDICTION IN MOOCS

Prediction in MOOCs is a trending area of research that aims to identify learners at risk and improve learners' learning experience. Predictions can also be useful to balance the cognitive load of course contents and improve the course design (e.g., sequence of materials, methodology, etc.) to reduce dropout and improve both engagement and learning outcomes. This topic has largely gained relevance in the past years, and many researchers have explored how to develop predictive models from different aspects, as shown by (blinded).

One of the aspects to consider regarding prediction in MOOCs is the variable to predict (also known as prediction outcome). Although the abovementioned review indicates that there can be many variables to predict, those related to learning outcomes (e.g., certificate earners, scores, and dropout) stand out, and particularly those related to learners' dropout. The former categories (certificate earners and scores) are more related to obtaining a passing grade and/or certification, while the latter (dropout) is more focused on completing the course, independently of the grade. On the one hand, among the former categories, Brinton & Chiang (2015), for example, predicted whether students were going to answer correctly their questions on their first attempt or not

(Correct on First Attempt, CFA). Ruipérez-Valiente et al. (2017) also compared different algorithms to predict certificate earners and found that boosted trees models were the most consistent ones when predicting using data from different weeks. On the other hand, regarding the latter category (dropout), Boyer & Veeramachaneni (2015), for example, forecast dropout in three different runs of the same MOOC and showed that models defined for the first two runs were accurately transferred to the third one.

Another aspect to consider when developing predictive models is the types of variables that can achieve better predictive power (i.e., predictor variables). The most common variables are those related to videos (e.g., Brinton & Chiang, 2015) and exercises interactions (e.g., Ruipérez-Valiente et al., 2017), and activity in the MOOC platform (e.g., Boyer & Veeramachaneni, 2015). For example, Xing & Du (2018) used variables about learners' activity (e.g., number of times learners accessed the course, checked the gradebooks, viewed the calendar, number of active days, etc.) to predict dropout over weeks in an 8-week MOOC. Furthermore, Hermans & Aivaloglou (2017) predicted course completion using many variables, including videos (e.g., number of videos watched, number of videos paused, total time spent watching videos, etc.) and exercises interactions (e.g., number of submitted questions, mean grade, mean submissions per question, etc.) from the first week.

Nevertheless, many other features which are not related to videos or exercises can be included in the predictive models. For example, Xing et al. (2016) included forum-related variables such as the number of posted messages per week and the number of forum views. Also, demographics and self-reported variables (e.g., familiarity with the topic, hours intended to be spent on the course) have already been studied (Greene, Oswald, & Pomerantz, 2015), and they can provide different information than event-based variables, which are supported by logs. This fact, together with the opportunities of the inclusion of new variables to analyze whether they can enhance current predictive models or not, leads to the analysis of new variables (SRL strategies) and the comparison between self-reported and event-based variables in RQ1 and RQ2.

However, the analysis of variables is not enough to have information about when is the best moment to predict. Regardless the potential of this analysis to inform about the indicators that have a better effect on learners' outcomes and/or behaviors, the predictive models only offer information about the learners' performance once the

course is finished, but not during the learning process. This fact, although often neglected, is very important to take corrective actions in the course. Thus, temporal analyses are needed. These temporal analyses have been already studied in prior work. For example, Kloft, Stiehler, Zheng, & Pinkwart (2014) analyzed dropout prediction over weeks using Support Vector Machines (SVM) and concluded that the predictive power was poor at the beginning of the course but it considerably improved at the end.

Currently, most of the contributions which consider the temporal analysis are conducted in instructor-paced MOOCs (synchronous). However, the same results might not apply for self-paced MOOCs, since learners participation in the course is asynchronous. The authors have already addressed these temporal analyses in instructor-paced settings in previous contributions (blinded, blinded). However, the scenarios under study were different because of three main factors: (1) the course methodology, (2) the predictive outcome, and (3) the available information. The MOOCs were instructor-paced in both articles and the predictive outcome was the assignment scores in (blinded) and the result (pass/fail) in an admission test in (blinded). Moreover, information about SRL strategies was not known in both cases, and little information about videos was considered in (blinded). Among all the differences, one of the most important ones is the course methodology. Because of that, and considering the aforementioned relevance of self-paced settings and the lack of research on them, this article aims to carry out the temporal analysis in a self-paced MOOC, as part of RQ3.

2.2. SELF-REGULATED LEARNING AND SUCCESS IN MOOCS

Recent studies have shown that SRL skills have a great impact on learners' success (Richardson, Abraham, & Bond, 2012), including dropout in MOOCs. Self-regulated learners are characterized by their ability to initiate cognitive, metacognitive, affective and motivational processes (Boekaerts, 1997). One of the problems in a MOOC is that learners often procrastinate their tasks, and thus, are more likely to drop out the MOOC (Michinov, Brunot, Le Bohec, Juhel, & Delaval, 2011). However, if they can set their goals, plan their work and reflect about their learning, i.e., self-regulate their process, they will be more likely to succeed (Wong, Baars, Davis, Van Der Zee, Houben, & Paas, 2019). The ability to regulate their own learning process is critical to achieve personal goals in a MOOC. Learners in self-paced MOOCs need to determine when and

how to engage with course content without any other support than the course content and structure, which can pose a challenge for many of them (Lajoie & Azevedo, 2006).

Prior research has shown the relationship between SRL skills and success. For instance, Broadbent (2017) found a positive relationship between SRL strategies and MOOC grades, and she highlighted the importance of time management and elaboration (i.e., ability to combine and build knowledge from multiple sources; Richardson et al., 2012) as two key SRL strategies. Hood et al. (2015) studied how SRL can affect learning outcomes and argued that the MOOC context played a crucial role on how learners self-regulate and engage in the course. Papamitsiou, Economides, Pappas, & Giannakos (2018) showed that most students with high time management skills perform better. Furthermore, Tempelaar, Rienties, & Nguyen (2018) concluded that students who only sporadically used examples with worked-out solutions achieved higher scores, which is related with the help seeking SRL strategy (i.e., tendency to ask for help when facing difficulties; Richardson et al., 2012).

Many researchers have also conducted empirical analyses about which specific variables and SRL strategies have an impact in dropout. For example, Y. Lee, Choi, & Kim (2013) explored SRL skills and found significant differences between metacognitive SRL skills (i.e., if students reflect on their understanding of materials) between students who dropped out and those who did not. Moreover, Sun, Xie, & Anderman (2018) found a significant positive relationship between self-efficacy (i.e., learners' beliefs about their capacities to achieve goals; Bandura, 1995) and help seeking SRL strategies and academic success. (Blinded) also found that time management and effort regulation were the most important SRL strategies for learners' goals attainment.

Despite previous articles have analyzed the relationship between SRL skills and success, the variables were usually self-reported (e.g., collected from surveys filled in by learners, as done by Sun et al., 2018) and not event-based. In addition, the analyses were usually carried out using statistical methods (e.g., correlations, as it is reported in several articles in the review of Lee et al., 2013), but not in predictive terms, which implies that our hypothesis has not been addressed yet. In a previous contribution, (blinded) started exploring the hypothesis through a regression analysis to predict learners' success in a MOOC. They used self-reported SRL strategies (which were also

related to actual behaviors in blinded) and sequence patterns of SRL (event-based), which were identified using process mining techniques (blinded). Results of this analysis gave an initial idea that event-based SRL strategies could achieve accurate results. Based on these previous contributions, a possible hypothesis is that event-based SRL strategies can be good predictors of dropout.

3. METHODOLOGY

An analytical methodology using data from a Coursera MOOC structured into two phases was followed in this article. The first phase consisted in the analysis of different aspects of dropout prediction in one MOOC. The second phase consisted in the validation of results in other two MOOCs. The aspects that will be considered in the analysis are: (a) the effect/importance of variables (and particularly, SRL strategies), (b) moment in which predictions are carried out, (c) algorithms, and (d) type of course. The first three aspects will be included in the analysis of the first phase while the last one will be part of the validation of the second phase. This section will introduce the data collection, measures, variables and analytical methods, including the algorithms and the metrics to evaluate them, to be used to conduct the study.

3.1. DATA COLLECTION

The first phase of the study was conducted in a MOOC on Electronics (named "Electrons in Action") offered by (blinded institution) in Coursera. The reasons to select the Coursera platform were mainly two: 1) because of availability reasons as this is the platform in which one of the institutions that are part of the article offer their MOOCs; 2) because the interactions on the Coursera platform allows the retrieval of data from which we can derive SRL strategies. Nevertheless, data from other platforms such as Open edX would also allow this detection of SRL strategies. Moreover, these data were used as they were gathered during the execution of a research project (blinded) and serves for the analysis of the Latin American region.

The MOOC was organized into four modules and contained 83 videos and 16 exercises (assessments). The passing rate was 80% and all assessments counted for the summative evaluation. The delivery language was Spanish, and the course was a self-paced MOOC. The MOOC was categorized as a xMOOC since the instructor is who provides

knowledge, and learners follow the coursework and ask questions where necessary (Kesim & Altınpulluk, 2015). The data collection period was between April and December of 2015. In this MOOC, there were 25,706 learners enrolled, although the sample size is N = 2,035 learners, as this was the number of learners who answered a self-reported online questionnaire about their SRL strategies. The data gathering techniques in this phase were the following.

1) Trace data from participants in the MOOC: They comprised Coursera logs and events (e.g., begin session, begin video lecture, complete video lecture, try assessment, complete assessment, etc.). They were used to obtain independent variables related to learners' activity, interactions with videos and exercises, and SRL sequence patterns. These data also served to retrieve the dependent variable, which is dropout, to train the predictive models.
2) Online questionnaire: It was an initial self-regulation questionnaire, which served as an instrument to define learners' SRL profile. It contained 35 questions, related to six SRL strategies (goal setting, strategic planning, self-evaluation, task strategies, elaboration and help seeking). This questionnaire also served to gather demographics information and learners' intentions (blinded). All the participants in the questionnaire filled a consent form revised by the ethical committee of the university to ensure privacy and correct use of data. This questionnaire was validated and showed high reliability with values of Cronbach's alpha of at least 0.70 for all subscales (more details in blinded)

A common limitation of predictive models is that it is not possible to ensure how these results can be transferred to other MOOCs because of the context, methodology, etc. In order to address this limitation and validate the results, two other MOOCs were evaluated in a second phase to analyze if similar conclusions could be obtained in other scenarios. These two MOOCs (also xMOOCs) were also offered by (blinded institution) in Coursera. They are both on Social Sciences, and their names are: "Constructivist Classroom" and "Management of Effective Organizations". They are also taught in Spanish in a self-paced mode, sharing these features with "Electrons in Actions". "Constructivist Classroom" comprises nine modules and 11 summative assessments (three of them were peer-review activities) while "Management of Effective Organizations" contains seven modules and six summative assessments. The data collection period was the same used in the first phase and the number of learners after

filtering (i.e., learners who completed the questionnaire) were 337 and 526 for both MOOCs, respectively.

3.2. MEASURES

Two measures were defined to conduct the analysis in both phases: (1) what dropout means in this study and (2) how time periods are defined to carry out the analysis in self-paced MOOCs. The second measure is important because in self-paced MOOCs learners may be engaging with different parts of the course in the same time period.

3.2.1. CONSIDERATIONS TO DEFINE DROPOUT

The predictive analysis of this paper is about forecasting dropout, which is the dependent variable in the analysis. This variable is categorical and classifies learners in two categories: "dropout" and "no dropout" (also referred as "completer"). While it is easier to define success in a MOOC (e.g., determine if a learner has achieved a passing grade), defining dropout can be more difficult since learners can be inactive for a period without dropping out of the course and continue after a while. This can happen particularly on self-paced MOOCs where learners do not have a defined schedule to complete the activities. Because of that, an analysis of how many learners interact after an inactive period was carried out. In order to do that, the inactivity period was computed as the maximum number of inactive days for a learner between two interactions. Percentiles of this variable are shown in Table 1.

Table 1

Percentiles of the inactive period of learners in the MOOC

Percentile	25	50	60	70	80	90	94	95	96	97	100
No. days	0	3	5	7	10	16	21	22	26	29	221

Results show, for example, that half of the learners show inactive periods of more than three days but then they interact again with the course. However, for the predictive purposes, it is interesting to find an inactivity period where learners are unlikely to interact again with the course. Therefore, an inactivity period of four weeks (28 days) was chosen for considering a dropout, as more than 96% of learners who are inactive for 28 days actually drop out the course (they never interact again).

Nevertheless, it is important to consider the period from the last interaction to the end of data collection. The reason is that a learner may have dropped out even when their period between interactions is short (e.g., if the learner interacts the first days, but does not appear anymore). Because of that, the inactivity period is defined as the maximum of the period between interactions and the inactivity period to the end of data collection. With this definition of inactivity period, two additional considerations were taken to define dropout:

1) As learners do not usually interact after they have completed the course, a learner was not considered as drop out if he/she submitted at least 80% of the assessments. In this case, grade will not matter because dropout is about competing the course (not passing it).
2) If a learner started the MOOC at the end of data collection period and/or slowly advanced in the course, it is not possible to label him/her as dropout or not. Because of that, if the inactivity period is below four weeks and the learner has not submitted at least 80% of the assessments, he/she will be discarded.

Under these considerations, a total of 926 learners were considered for the analysis. To summarize, Fig. 1 depicts a flow diagram with the rules to determine dropout.

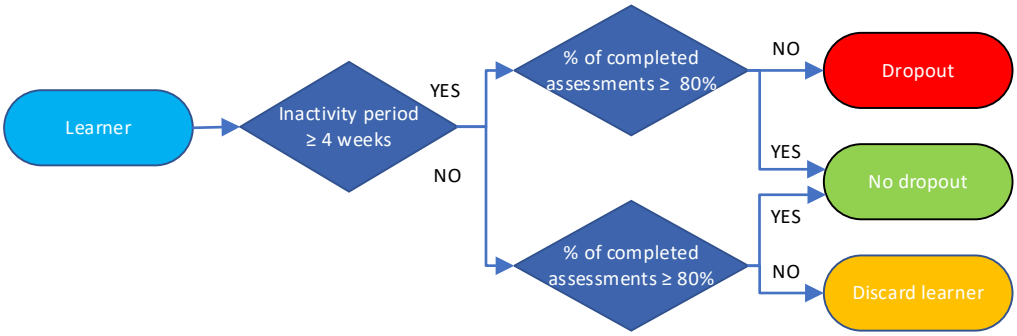

Fig 1 Flow diagram with the rules to determine dropout

3.2.2. CONSIDERATIONS RELATED TO THE SELF-PACED DELIVERY MODE

After defining the dependent variable of the problem, we stablished the criteria for conducting the temporal analysis. In an instructor-paced MOOC, temporal analyses often consider the week as the unit of time and analyze how the predictive power varies when adding data of consecutive weeks (e.g., Okubo, Yamashita, Shimada, & Ogata, 2017). Nevertheless, this cannot be transferrable to self-paced MOOCs because learners

can be engaging with different parts of the MOOC at the same time (e.g., week). In the literature, solutions to this problem are scarce. Vitiello, Walk, Helic, Chang, & Guetl (2018) had a similar problem and they considered the interactions within the first 1% to 100% of the total active time, so interactions were normalized with the active time. Although this can be reasonable, low variance between features was detected when the level of course completion was similar among learners. Instead, in this article, another solution is proposed, which is the normalization of time period (instead of active time), as it was seemed to be more representative.

This normalization consists on computing the period from the first interaction of each learner to the end of the selected period (e.g., if one week is considered, the period would comprise the seven first days from the first interaction) and perform the analysis using weeks as a unit of time (as it is usually done in instructor-based MOOCs). Fig. 2 illustrates this normalization. In this case, the analysis of the first week will be the first block (week) of each learner, the analysis of the first two weeks will consider the first two blocks, and so forth.

Although this approach does not consider the speed at which learners advance in the course, data showed that learners who completed the MOOC typically completed the modules in the suggested period (as the course had four modules, it was suggested doing it in four weeks), although they might be faster or slower at certain times. Particularly, data showed that 48% of the interactions occurred in the first week, 69% in the first two weeks and 94% of them in the first four weeks. This makes using the abovementioned normalization reasonable, as the it will capture the information about when learners interacted with the platform respect to their first activity.

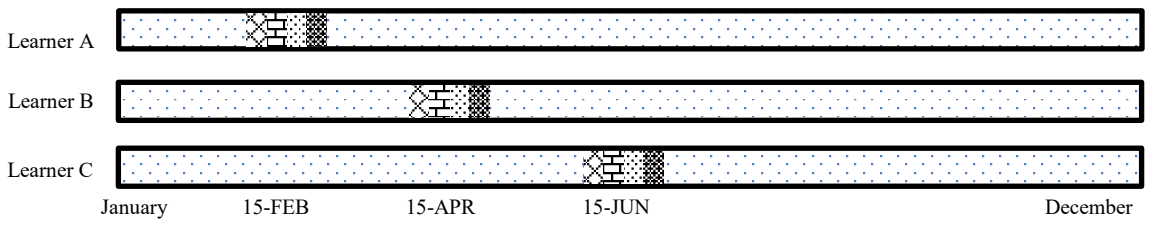

Fig 2 Example of the first four weeks (blocks) of the course for three learners starting the course at different time periods

3.3. Analytical Methods and Variables

Once data are collected and measures and defined, the last step is obtaining the indicators that will be used for prediction. From the data sources mentioned in Section 3.1, several independent variables were obtained. In order to facilitate the analysis and understand better the importance of these variables, seven categories were defined: (1) self-reported SRL strategies, (2) SRL sequence patterns (event-based), (3) demographics, (4) intentions, (5) activity, (6) videos and (7) exercises. Categories (1), (3) and (4) were obtained from the first mentioned source of data (questionnaire), while the rest were obtained from the second source (Coursera logs and events). Table 2 contains the full list of variables and their descriptions. In RQ1 and RQ2, dropout predictive models were developed for each category using all the available interactions, together with a model with all categories and combinations with SRL strategies.

Table 2

List of variables used in the study

Variable	Description
Variables related to self-reported SRL Strategies	
(1a) GoalSetting	Value coded in range 0-4 about goal setting strategies
(1b) StrategicPlanning	Value coded in range 0-4 about strategic planning strategies
(1c) SelfEvaluation	Value coded in range 0-4 about self-evaluation strategies
(1d) TaskStrategies	Value coded in range 0-4 about task strategies
(1e) Elaboration	Value coded in range 0-4 about elaboration strategies
(1f) HelpSeeking	Value coded in range 0-4 about help seeking strategies
Variables related to SRL Sequence patterns (event-based)	
(2a) Only_vlecture	Pattern identified when learners only interact with videos
(2b) Atry_to_Vlecture	Pattern identified when the learner opens the assessment first and then looks for information in videos
(2c) Explore	Pattern identified when learners superficially inspect the contents without intention to complete them
(2d) Only_assessment	Pattern identified when learners only interact with assessments
(2e) Vlcomplete_to_Atry	Pattern identified when learners interact with videos and then start an assessment (without finishing it)
(2f) VLecture_to_Acompl	Pattern identified when the learners interact with videos and then start and assessment (finishing it)
(2g) Complex	Variable for other patterns not in included in (2a)-(2f)
Variables related to demographics	
(3a) Edu	Educational level
(3b) Age	Age of the learner
(3c) Isfemale	Categorical variable representing whether the learner is male or female
(3d) Emp_student	Categorical variable representing whether the learner is a student (in formal education) or not
(3e) Emp_job	Categorical variable representing whether the learner has a job or not
Variables related to learners' intentions	
(4a) Hrs	Number of hours the learner intends to dedicate to the course
(4b) Int_topic	Categorical variable representing whether the learner is interested in the MOOC topic or not
(4c) Int_assess	Categorical variable representing whether the learner intends solving the assessments or not
(4d) Prior_exp	Categorical variable representing whether the learner has previous experience with MOOCs or not
Variables related to learners' activity	
(5a) Days_Act	Number of active days in the platform
(5b) Time_spent_min	Total time spent interacting in the platform (in minutes)
(5c) Num_ses	Number of sessions
Variables related to learners' interactions with videos	
(6a) Vl_complete	Number of times the learner has completed a video
(6b) Vl_begin	Number of times the learner has started watching a video without finishing it
(6c) Vl_review	Number of times the learner has reviewed a video once completed
(6d) Prop_vlopen	Percentage of opened videos (completed or not)
(6e) Prop_vlcomplete	Percentage of completed videos
(6f) Prop_vlreview	Percentage of reviewed videos
Variables related to learners' interactions with exercises	
(7a) A_try	Number of times the learner has started to do an assessment without finishing it
(7b) A_complete	Number of times the learner has completed an assessment
(7c) A_review	Number of times the learner has reviewed an assessment once previously completed successfully
(7d) Prop_atry	Percentage of attempted assessments (completed or not)
(7e) Prop_acomplete	Percentage of completed assessments

Among SRL strategies, it is important to note that category (1) is referred to self-reported SRL strategies (i.e., obtained from a questionnaire), while category (2) is related to event-based SRL sequence patterns, which were obtained from Coursera logs. In order to obtain those patterns, process mining techniques following the PM^2 method were used (blinded), together with agglomerative hierarchical clustering based on Ward's method. As a result, six main patterns were identified (2a) to (2f) in Table 2. Feature 2g was reserved for other patterns which were not classified as the main ones. More details about the extraction of these variables, and the justification of the reliability of the patterns can be found in a previous contribution (blinded).

Four classical machine learning algorithms were used as predictive algorithms: (1) Random Forest (RF), (2) Generalized Linear Model (GLM), (3) Support Vector Machines (SVM) and (4) Decision Trees (DT). For the implementation of the predictive models using these algorithms, caret[1] package of R was used. In addition, results are obtained using 10-fold cross validation to obtain a higher reliability of the results with a good compromise between bias and variance (McLachlan, Do, & Ambroise, 2005); AUC (Area Under the Curve) is used as a metric to evaluate the predictive models. The reason for using AUC is that it has been shown appropriate for this kind of classification problem (Pelánek, 2015) (i.e., dropout prediction) and results are not biased depending on the dataset balance (Jeni, Cohn, & De La Torre, 2013) (e.g., accuracy, for example, can be high even for a poor model if the dataset is imbalanced).

4. RESULTS AND DISCUSSION

Next, each subsection details the analysis of the research question and the discussion of the results, according to the two phases of the methodology (analysis in one MOOC and validation in other two MOOCs) described in Section 3.

4.1. RQ1: WHAT IS THE PREDICTIVE POWER OF SELF-REPORTED SRL STRATEGIES IN DROPOUT PREDICTION

[1] http://topepo.github.io/caret/index.html

The first part of the analysis aimed to analyze the effect of self-reported SRL strategies when predicting dropout. For that, dropout predictive models were developed for the different categories of variables. Results of the predictive models (expressed in AUC) can be found on Table 3. This table contains models not only related to self-reported SRL strategies, but also related to the seven categories. Therefore, this table can be used to compare different sets of variables. Best model for each set of features is highlighted in italics, best model for each algorithm is marked in bold and the best overall model has an asterisk.

Table 3

Results of predictive models using different sets of features (expressed in AUC)

Features	**DT**	**GLM**	**RF**	**SVM**
Self-reported SRL strategies	0.51	*0.57*	0.52	0.49
SRL sequence patterns (event-based)	0.85	***0.96***	*0.96*	0.94
Demographics	0.55	*0.63*	0.57	0.57
Intentions	0.50	*0.59*	0.48	0.54
Activity	0.86	*0.94*	0.92	0.89
Activity and SRL Sequence Patterns	0.89	***0.96***	*0.96*	0.94
Videos	0.91	*0.93*	*0.93*	0.92
Videos and SRL Sequence Patterns	0.92	0.95	*0.96*	0.94
Exercises	***0.96***	***0.96***	*0.96*	***0.96***
Exercises and SRL Sequence Patterns	**0.96**	**0.96**	***0.97****	0.95
All	***0.96***	***0.96***	*0.96*	0.93

An initial observation is that the predictive power of self-reported SRL strategies is the lowest among all the sets of features and it is significantly worse than the rest of models. This indicates that variables about self-reported SRL strategies are useless for prediction. A possible explanation is that self-reported SRL strategies were obtained from a questionnaire where learners had to indicate how a set of statements (e.g., "*When I am learning, I try to relate new information I find to what I already know*") described their behavior in range 1-5 (value coded between 0-4 in Table 2), and answers may have not reflected the actual learner behavior, e.g. learners might have not been aware of their self-reflection abilities or some of them might have lied. Moreover, prior literature indicates that these self-reported measures are limited when taken at the beginning of the course, where students have expectations about their behavior of the course without knowing it (Panadero, Klug, & Järvelä, 2016).

For example, it is typical to ask learners how many activities they plan to do and it may happen (as it occurred in Engle, Mankoff, & Carbrey, 2015) that most of the learners believe that they are going to complete all or most of them (78.2% in Engle et al., 2015), although they may actually not take any assignment. This fact is related to the reliability of self-reported data (as discussed in Panadero et al., 2016), which can be limited. The reason is that data can be biased because of learners' beliefs and motivation, and that can make self-reported SRL strategies have low predictive power. Similar arguments can be given to justify why the predictive power using variables related to intentions is low. Regarding demographic features, these also achieve poor results. This result is in line with those of Brooks, Thompson, & Teasley (2015), where demographic variables offered minimal performance with respect to activity features. The last fact about activity features is also confirmed with these data as event-based variables, related to users' activity and interactions with videos and exercises, can indeed achieve powerful predictive results.

For the second phase of the methodology (validation), results of the predictive models (using the same criteria as in the first phase) are presented in Table 4. With regard to the self-reported SRL strategies, it can be seen that the predictive power is not any better than in "Electrons in Actions". A similar poor performance is obtained with the other sets of features which are not event-based (demographics and intentions), which confirms that they are useless for prediction purposes. In contrast, variables obtained from Coursera logs, such as activity, videos and exercises, also achieve powerful results, recommending their use for the predictive models.

Table 4

Results of predictive models using different sets of features (expressed in AUC)

Course	**Constructivist classroom**				**Management of Effective Organizations**			
Features	**DT**	**GLM**	**RF**	**SVM**	**DT**	**GLM**	**RF**	**SVM**
Self-reported SRL Strategies	0.54	0.61	*0.62*	*0.62*	0.50	*0.55*	0.46	*0.55*
SRL Sequence patterns (event-based)	0.88	*0.97*	0.95	0.94	0.75	0.94	0.95	*0.96*
Demographics	0.54	0.46	*0.57*	0.55	0.51	*0.55*	0.50	0.52
Intentions	0.50	*0.62*	0.53	0.58	0.50	*0.56*	0.50	0.50
Activity	0.92	0.93	0.95	*0.96*	0.87	*0.94*	*0.94*	*0.94*
Activity and SRL Sequence Patterns	0.92	0.96	*0.97*	0.94	0.87	*0.96*	*0.96*	*0.96*
Videos	0.92	*0.97*	*0.97*	*0.97*	0.92	0.94	*0.97*	*0.97*
Videos and SRL Sequence Patterns	0.92	*0.97*	*0.97*	0.93	0.92	0.95	*0.97*	*0.97*
Exercises	**0.97**	**0.98**	***0.99****	**0.98**	***0.97***	***0.97***	*0.97*	*0.97*
Exercises and SRL Sequence Patterns	**0.97**	***0.98***	*0.98*	0.96	**0.97**	0.95	0.97	***0.98***
All	**0.97**	0.89	***0.99****	0.95	**0.97**	0.95	***0.99****	0.94

4.2. RQ2: WHAT IS THE PREDICTIVE POWER OF EVENT-BASED SRL STRATEGIES IN DROPOUT PREDICTION

In this section, an analysis of whether results vary when including event-based SRL strategies instead of self-reported ones is carried out. In order to address this issue, results of the predictive models using event-based SRL sequence patterns, which are presented in Table 3, are discussed. The predictive power of SRL sequence patterns reveals that event-based SRL strategies can achieve an AUC up to 0.96 with RF by themselves, which is a good result and considerably better than results obtained with self-reported SRL strategies. This finding supports the idea that SRL skills in a MOOC can affect learning outcomes and particularly dropout, as SRL sequence patterns achieve powerful performances without the need of other variables.

Despite SRL sequence patterns can be useful to predict, we analyzed their predictive power compared to other event-based features (activity, videos and exercises) and whether or not SRL sequence patterns can improve their predictive power for dropouts when mixed with other indicators. With regard to activity features, results show that they can also achieve a strong predictive power (up to 0.94), but results are generally worse than SRL sequence patterns (except with DTs). Moreover, the predictive power slightly improves when adding SRL sequence patterns in all the algorithms (the AUC improvement is between 0.02 to 0.05). Similar results are obtained with video features, where SRL variables tend to be better (except with DTs) and results also slightly improve when mixing both kind of features. However, it can also be observed that, when analyzing variables related to exercises, these variables have the same predictive power as SRL patterns (and even better AUCs with DTs and SVMs). Furthermore, results show that the addition of both SRL patterns and exercises does not have any significant effect. A possible reason is that SRL patterns are related to interactions with exercises because SRL patterns are obtained from events related to interactions with assignments. AUC slightly improves up to 0.97 with RF but results barely change in general. A similar effect is found when combining all variables. In that case, results are the same than those obtained with variables related to learners' interactions with exercises, except for AUC with SVM, which is lower, perhaps due to the possible noise of some variables.

The last result suggests that variables related to learners' interactions with exercises are the best predictors. Actually, they predict well enough without the need to add SRL strategies. Nevertheless, SRL sequence patterns achieve a good performance by themselves, which entails that they can be useful for prediction. For example, in this case, they would be useful provided that exercises logs were not available. Moreover, the predictive power when mixing SRL sequence patterns with videos or activity slightly improved, which highlights the importance of SRL. The slightly improvement may be explained because there might be relationships between variables, and learners who show high values in the SRL patterns variables are more likely to be engaging in the course and achieve higher level of interactions with videos and exercises.

In order to provide more evidences about which variables have stronger effect on the predictive models, we evaluated the importance of variables using the Mean Decrease Gini, a common measure for the importance of variables based on the node impurity (Louppe, Wehenkel, Sutera, & Geurts, 2013). The model used to compute the variable importance uses all the variables at the end of the course (the last row, named "All", in Table 3) and RF. Table 5 indicates the variable importance for the indicators used in the study, ordered from higher to lower importance.

Table 5

Variable Importance (VI) of features using Mean Decrease Gini

Variable	VI	Variable	VI	Variable	VI	Variable	VI
Prop_acomplete	32.07	Num_ses	12.12	Age	3.03	Complex	1.42
A_complete	30.88	Only_assessment	8.39	Vl_begin	2.40	SelfEvaluation	1.32
Prop_atry	28.03	Only_vlecture	7.39	HelpSeeking	2.03	Vlecture_to_Acompl	1.29
Vl_complete	25.22	Explore	7.32	StrategicPlanning	1.99	Vlcomplete_to_Atry	1.17
Prop_vlopen	24.70	Prop_vlreview	5.72	GoalSetting	1.98	Prior_Exp	0.99
Prop_vlcomplete	21.78	A_review	5.65	TaskStrategies	1.75	Emp_student	0.69
Time_spent_min	17.89	A_try	4.16	Hrs	1.69	Emp_job	0.56
Days_act	15.40	Vl_review	3.67	Edu	1.63	Int_Assess	0.51
Atry_to_vlecture	14.15	Prop_areview	3.61	Elaboration	1.51	Isfemale	0.38

A first observation shows that the most important variable is the percentage of completed assignments. This is very reasonable at the end of data collection period because of the relationship between assignment completion and dropout (i.e., learners need to complete the assessments to complete the course) and it can also explain why variable related to exercises behave so well. In this case, it is important to note that although assessment completion is highly related to dropout, it should not be removed to evaluate the models because, despite it can indicate dropout at later stages, there is no

direct relationship at early stages and it can be very useful to anticipate what will happen in the MOOC.

After variables related to learners' interactions with exercises, the most important ones are those related to the number of videos learners open and the number of videos learners complete. This entails that it is very important to watch the video lectures to complete the MOOC and although there may be learners who skip some videos because they already master part of the topics, they need to cover most of them to grasp the contents. In contrast, video reviews seem to have lower predictive power because although they are more frequent within the group of completers, not many learners review their videos. Other activity variables (i.e., time spent and days active) are also among the best predictors.

With regard to the SRL sequence patterns, the most important feature is *Atry_to_vlecture*, followed by *only_assessment*, *only_vlecture*, and *explore*. The remaining patterns (*Vlecture_to_Acompl*, *Vlcomplete_to_Atry*) do not have a high variable importance in the model. Fig. 3 presents boxplots between sequence patterns and dropout. For the variable with higher importance (*Atry_to_vlecture*), the figure shows that the first quartile of this variable for those who finish the course is almost the same as the maximum value for those who drop out the MOOC. Therefore, this variable can be useful to classify learners because of the high differences between the values for the two classification classes (learners who drop out or not). In contrast, there are very few cases for the pattern *VLecture_to_Acomplete* and *Vlcomplete_to_Atry* for learners who drop out the MOOC, and for those who complete it. This implies that, although the instructor normally suggests watching the video lectures first and then attempting the assessments, this pattern is not very common. This is in contrast to the opposite pattern, where learners open the assessments first and they look for the answers in the videos (which is in fact the most common pattern). The consequences of this fact are that patterns which start with interactions with videos to then continue with interactions with assessments have low predictive power, as it is shown in Table 5.

Patterns where learners only interact with videos or assessments also present differences between users who drop out and with those who do not. It is interesting though that the prominent pattern for those who drop out the course is *only_vlecture*, which might be because they can "sample" some videos or they enroll to focus on specific content that

helps them to meet goals elsewhere (lurkers and drop-ins profiles, respectively, according to Hill, 2013). After this pattern, it is common that dropout learners explore the course (e.g., open a lecture and assessment without completing any of them), although this is also typical for completers.

Finally, it is noteworthy that variables obtained from demographics and from the questionnaire have low effect on prediction. In fact, variables such as the occupation of the learner, interest of topic and gender are the variables with the lowest variable importance in Table 5. Data about interest of topic is a clear indicator of the differences between what learners say in the questionnaire and what occurs in reality. In the MOOC, 91.5% of learners indicated that they were interested in doing the assignments despite 79.7% dropped out. Variables related to occupation and gender show that the demographic variables are not good predictors of dropout, as it also concluded in Willging & Johnson (2009).

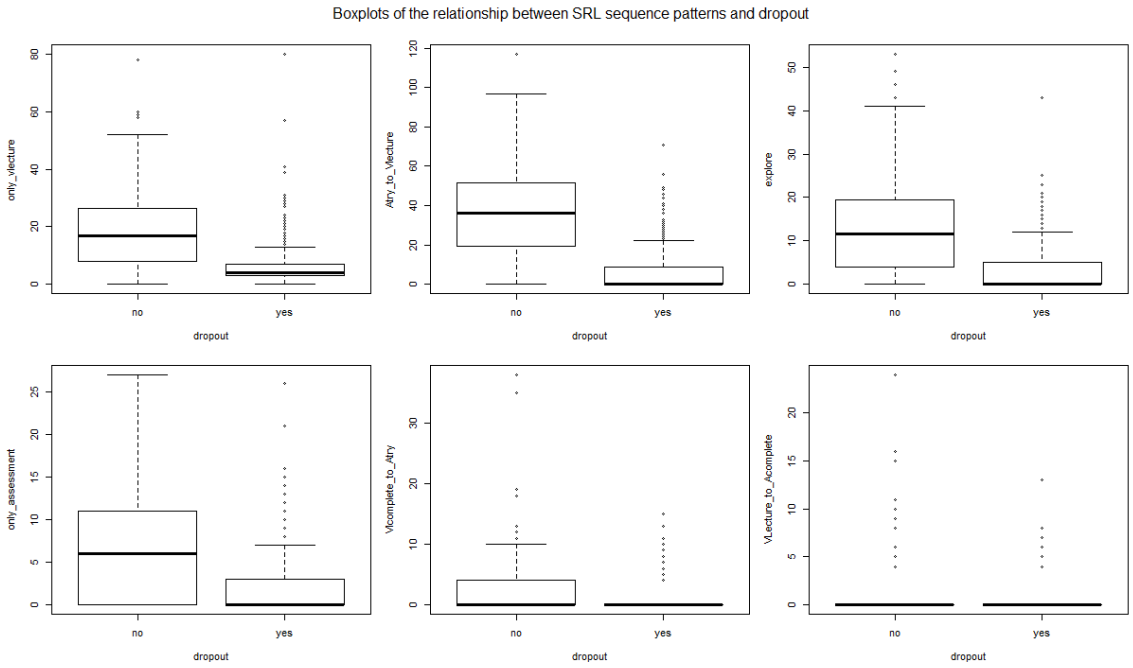

Fig 3 Boxplots of the relationship between SRL sequence patterns and dropout

After analyzing the importance of variables and particularly event-based SRL strategies in one MOOC, we have validated the results using the other two MOOCs. In this case, results (see Table 4) show that SRL sequence patterns are also good predictors (their AUC can be up to 0.97 in Constructivist Classroom and up to 0.96 in Management of Effective Organizations). Furthermore, it can be seen that the predictive power of

activity features generally slightly improves when mixing those variables with sequence patterns as well. However, this does not apply to variables related to learners' interactions with videos in these two MOOCs, unlike in Electrons in Actions. Variables related to learners' interactions with exercises also stand out and they seem to predict well enough, at least at later stages. Nevertheless, the predictive power with all variables is outstanding, particularly with RF where AUC is 0.99.

These results, in relation to those presented in the first MOOC, clarify that **the best predictors are the variables related to learners' interactions with exercises, and they can be enough to get a good prediction.** These results match with a finding of a previous contribution in which previous grades had a considerable predictive power to predict future grades (blinded). Nevertheless, the **predictive power of SRL sequence patterns is also good in all cases, which suggests that this kind of variables can be used in other contexts alone or together with other kind of variables if they are available**. However, it is important to note that the effect of the features may vary depending on the course and it is possible that variables related to learners' interactions with exercises can be complemented better with other variables in other contexts. Because of that, further research can be done to explore the effect of SRL strategies.

As a final experiment, an evaluation was conducted to test how transferrable models are when applied to different MOOCs. This evaluation is relevant to gather conclusions about how generalizable predictive models and the findings of this work are. In order to do that, models using all the available data have been trained with each of the three MOOCs and they have been tested with the data of the other two MOOCs. Results of this analysis can be found in Table 6.

Results show that it is possible to achieve a high AUC (over 0.96) when using data from different MOOCs for training and testing the predictive models. This means that it is possible to reuse predictive models regardless the duration and the thematic area of the course in this context. However, results show that DT and GLM are not always consistent, and they offer models not transferrable when training with Electrons in Action and Constructivist classroom. This leads **RF to be the best model as it is the most consistent when transferring to other MOOCs**. Moreover, it is the algorithm which achieves higher predictive power, both when testing in the same MOOC or in a different one.

Table 6

Results of transference between predictive models among courses (expressed in AUC)

Train course	Electrons in Action (EIA)		Constructivist classroom (CC)		Management of Effective Organizations (MEO)	
Test course	**CC**	**MEO**	**EIA**	**MEO**	**EIA**	**CC**
DT	0.50	0.50	0.94	0.50	0.89	0.96
GLM	0.95	0.54	0.28	0.75	0.96	0.96
RF	0.99	0.97	0.97	0.96	0.97	0.98
SVM	0.95	0.96	0.91	0.94	0.89	0.92

Despite the predictive models and the findings of this work are transferrable to the other two MOOCs, it is noteworthy that this not always necessarily happens and particularly when changing to more different contexts. In this case, at least the three courses where self-paced MOOCs taught in Spanish with similar assessment criteria, although their duration and thematic area were different. If the courses had been even more different, results may have differed, and particularly if models used data from different sources. For example, if each course was hosted in a different platform, and the information from the traces were not equivalent, the list of indicators could vary depending on the platform. Consequently, the analysis and the results may differ as well. This raises the importance of considering the MOOC context to reach the generalizability by the adaptation of the predictive models whenever the context requires to do so.

4.3. RQ3: WHEN IS THE BEST MOMENT TO PREDICT DROPOUT IN A SELF-PACED MOOC?

In previous sections, models have always been developed using data obtained after the end of the course. This can be useful for a posteriori analysis and for detecting patterns, but it is not practical. The reason is that if predictions are made at the end of the course, there will not be time to react and adapt teaching/learning behaviors (blinded) to run interventions. Because of that, it is interesting to discover how much it is possible to predict in advance with accurate results. Obviously, there will be a trade-off between anticipation and predictive power (i.e., if you predict earlier, your predictive power will be worse as there is less information), but if results are good enough at early stages, they are preferable.

In order to address this issue, the performance of predictive models over the time has been evaluated. Models that only use exercises and SRL sequence patterns have been chosen as their predictive power seemed to be better in this context (as shown in Table 3), and the four aforementioned algorithms have been used for the analysis. As for the

timing, Fig. 4 shows the evolution of the predictive power for the different models using data from the interactions of the first five weeks of each learner in the MOOC (one model with only the first week, another with the first two weeks and so on).

Results show that it is possible to predict with a good AUC (between 0.8 and 0.9, according to Mezaour, 2005) and with and excellent AUC (greater than 0.9) from the second week. This finding shows that it is possible to predict whether a learner is going to drop the course with very few days from the first interaction. It also means that the first days are very representative as they can be used to predict with high predictive power. This result matches with the result that Jiang, Williams, Schenke, Warschauer, & O'dowd (2014) obtained when they predicted certificate earners in a MOOC with only interactions with the first week. In their contribution, their course was synchronous (i.e., instructor-based), but results here indicate their finding can be extrapolated to asynchronous self-paced environments. Nonetheless, the figure also confirms that the predictive power increases over time and reaches a peak at 0.97. In a live course, this means that predictions could also be improved and updated over time to provide instructors and/or learners the best information (e.g., alerts) about learners at risk.

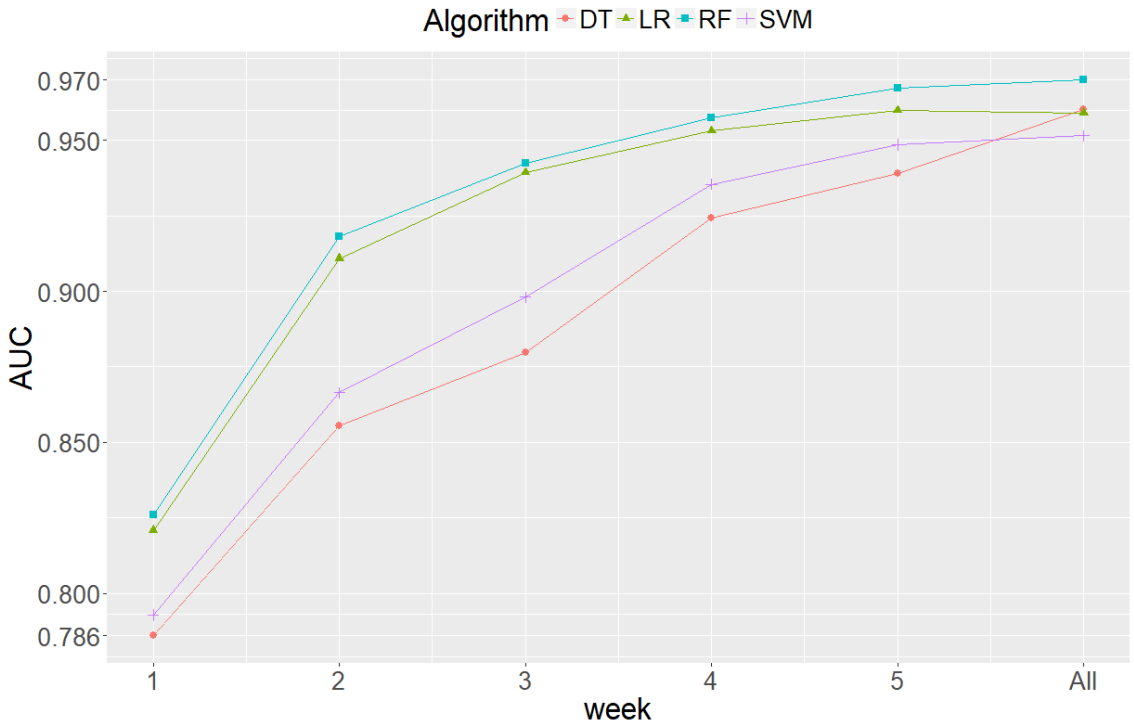

Fig 4 Evaluation of results of dropout prediction over time

With regard to the algorithms, Fig. 4 illustrates that RF is the best algorithm among the four used in the analysis in all the selected periods. However, the difference with LR is small, particularly in the first weeks. In contrast, DTs and SVM seem to achieve worse performance during all the weeks despite their predictive power can be considered acceptable as well. This result entails that RF should be preferred as it outperforms the rest of the algorithms, although focus should be also put on the features because they can also make the difference, as demonstrated in Section 4.1.

As a validation of these findings (second phase of the methodology), a comparison of the results with the other two MOOCs has been conducted. This will be particularly interesting since the "ideal" duration of each MOOC is different (the number of modules of each MOOC varies and thus the number of weeks that each learner should work to finish the course). In this case, as the longest MOOC comprises nine modules, models with the first ten weeks of each learner will be obtained as well as a model with all the interactions within the data collection period. The model with all the variables has been considered because of its high predictive power in all the scenarios. Fig. 5 depicts the results of the predictive power over time.

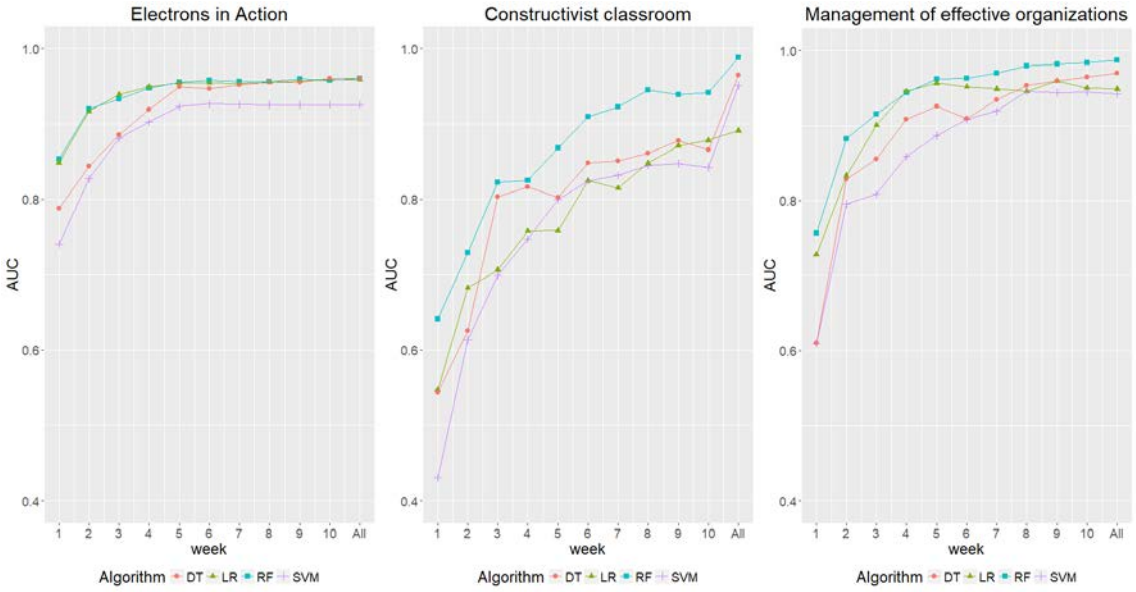

Fig 5 Evaluation of results of dropout prediction over time in the three MOOCs

An initial observation is that the predictive power reaches a steady state in Electrons in Action after week 6. The reason is that 99% interactions occur within that period, and

therefore there is barely new information after week 6. This also implies that learners who do not drop out try to follow the suggested schedule (although they may spend a couple of weeks more than expected to complete the activities). A similar effect occurs at about week 10 in Management of Effective Organizations (which is 3 weeks more than the theoretical duration). At that point, 99.6% of the interactions have occurred. However, the case is different for the Constructivist Classroom, which is the longest MOOC and only 89.7% of the interactions have occurred at week 10.

As for the predictive power, taking the threshold of 0.8 for considering the AUC as good, the MOOCs Constructivist Classroom and Management of Effective Organizations reach that point at weeks 3 and 2, respectively. The threshold of 0.9 (excellent AUC) is reached at weeks 6 and 3, though. This means that between 25-33% of the total theoretical MOOC duration is needed for an AUC of 0.8 and between 43-67% for an AUC of 0.9 considering all the MOOCs. This finding will be particularly useful because it allows developing early predictive models that can be used to alert learners at risk and produce a positive effect on their learning.

5. CONCLUSIONS

This work presents the results of analyzing different predictive models with a variety of features (including novel features related to SRL) and algorithms at several time periods. Results show that variables related to learners' interactions with exercises are the best predictors, as happened in other contributions (Ruipérez-Valiente et al., 2017; blinded). However, features related to videos and SRL sequence patterns (event-based) also achieve a high predictive power by themselves. These results suggest that SRL sequence patterns can be included if available, particularly if variables related to learners' interactions with exercises are not available or scarce (e.g., in a MOOC where exercises are released gradually and interactions with exercises can be limited at certain stages of the course). In contrast, variables obtained from questionnaires (including self-reported SRL strategies) and demographics show not to be useful for prediction. This result matches with findings by Brooks et al. (2015), although they did not explore SRL strategies. Therefore, the answers to RQ1 and RQ2, respectively, are that self-reported SRL variables achieve poor predictive power, while event-based SRL variables can be strong predictors, although worse than variables related to exercises. Nevertheless, it

could be relevant to analyze more SRL patterns to analyze whether alternative results could be obtained. Furthermore, the fact that self-reported SRL variables are not good predictors also raises some alternative hypothesis for further research: 1) learners are not aware of their SRL skills, 2) learners do not take the questionnaire seriously (as mentioned by Panadero et al., 2016) and some changes should be done to handle this issue, and 3) another survey may be needed.

In relation to the best predictors, those related to the number and percentage of assessments opened and completed stand out, followed by those related to videos opened and completed. Among the SRL sequence patterns, the more predictive ones are the patterns of learners who start assessments and then go to video lectures to look for questions, and those which only contain videos or exercises interactions.

With regard to the best moment to predict (related to RQ3), the analysis shows that 25-33% of the theoretical duration of the MOOC (considering one week per module) is enough time to predict with good predictive power. We also show that the approach of considering the weeks from learners' first interaction with the MOOC is also useful to predict in a self-paced MOOC. This happens because learners' interactions mainly happened in a period which was similar to the theoretical duration of the MOOC. Moreover, the validation of results shows that predictive models are transferrable to other two MOOCs, as the predictive power when using models trained with one course and tested with other is very high (up to 0.99 of AUC).

The abovementioned findings, in relation to existing related research work, advances current predictive models with the following contributions:

C1) Inclusion of new variables about SRL that have not been considered in the literature for dropout prediction, and their analysis of whether they can improve the predictive power or not and whether they can achieve powerful results by themselves or not (addressed in both RQ1 and RQ2)

C2) Analysis and comparison between self-reported SRL strategies and event-based SRL strategies for dropout prediction (addressed in RQ1 and RQ2)

C3) Temporal analysis to analyze dropout prediction in a self-paced MOOC (addressed in RQ3), with a specific and novel methodology for self-paced MOOCs (addressed in Section 3.2)

Particularly, previous contributions only focused on data about clickstream (e.g., Halawa et al., 2014; Ye, et al., 2015). While this approach can be valid, and high predictive power can be achieved (e.g., Laveti et al., 2017), it is also relevant to examine SRL behaviors from MOOC data, as indicated by (D. Lee, Watson, & Watson, 2019), because it is found that SRL variables can also achieve high predictive power by themselves. This is a novel contribution (contribution C1) and provides insight about new variables to be used in the predictive models. Furthermore, SRL has been mainly measured through self-reported variables (e.g., Hood et al., 2015) and this paper incorporates event-based variables, which reflect actual behaviors, and compares them with traditional self-reported variables in terms of prediction (contribution C2). In addition, this paper analyzes the best moment to predict in a self-paced MOOC. Other contributions have taken similar approaches in instructor-paced MOOCs, but this analysis is also important because time periods vary for each learner in self-paced MOOCs (contribution C3).

Despite the abovementioned findings, there are some limitations to mention. First, data were limited and it only comprised a limited period of time. Although models behaved well, more data may help to improve their predictive power. This period of time also limited the number of learners whose classification (completer or dropout) was known, and it was needed to filter out the learners whose classification could not be determined. Furthermore, the measures (dropout and self-paced considerations), although they were justified for this scenario, they are not unique, and other measures may modify the results. In addition, SRL was measured through a questionnaire and sequence patterns, but these sequence patterns can be related to other features such as videos or exercises (e.g., if a learner watch video lectures after opening assessments many times, it indirectly means that he is interacting with videos and exercises)

As future work, we plan to explore other ways to measure SRL strategies through logs (not only questionnaires), as current results show that they can potentially be useful for prediction. Moreover, it would be interesting to analyze other self-paced settings (not MOOCs) to analyze how findings are applicable to other different scenarios. Furthermore, next step should focus on incorporating these predictive models into a tool to give insights to teachers and/or learners and promote with timely interventions. As prediction models seems to be transferrable, at least with the analyzed courses, models should be used in current sessions of those MOOCs (which are currently offered on

Coursera) and pilots should be developed to ensure they are useful within the course context and they have a practical effect on education.

REFERENCES

Bandura, A. (1995). Comments on the crusade against the causal efficacy of human thought. *Journal of Behavior Therapy and Experimental Psychiatry, 26*(3), 179-190.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and instruction, 7*(2), 161-186.

Bote-Lorenzo, M. L., & Gómez-Sánchez, E. (2017, March). Predicting the decrease of engagement indicators in a MOOC. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference* (pp. 143-147). ACM.

Boyer, S., & Veeramachaneni, K. (2015, June). Transfer learning for predictive models in massive open online courses. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (pp. 54-63). Springer, Cham.

Brinton, C. G., & Chiang, M. (2015, April). MOOC performance prediction via clickstream data and social learning networks. In *Proceedings of the 2015 IEEE Conference on Computer Communications* (pp. 2299-2307). IEEE.

Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education, 33*, 24-32.

Brooks, C., Thompson, C., & Teasley, S. (2015, March). Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs). In *Proceedings of the 2nd (2015) ACM Conference on Learning@ Scale* (pp. 245-248). ACM.

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*.

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of interactive Media in education, 2012*(3), Art-18.

Engle, D., Mankoff, C., & Carbrey, J. (2015). Coursera's introductory human physiology course: Factors that characterize successful completion of a MOOC. *The International Review of Research in Open and Distributed Learning, 16*(2).

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding Dropouts in MOOCs. In *Proceedings of the 33rd AAAAI Conference on Artificial Intelligence* (In Press). AAAI.

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of retention and achievement in a massive open online course. *American Educational Research Journal, 52*(5), 925-955.

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. In *Proceedings of the 2nd European MOOC Stakeholder Summit* (pp. 58-65).

Hermans, F., & Aivaloglou, E. (2017, May). Teaching software engineering principles to k-12 students: a mooc on scratch. In *Proceedings of the 39th International Conference on Software Engineering: Software Engineering and Education Track* (pp. 13-22). IEEE Press.

Hill, P. (2013). Emerging student patterns in MOOCs: A (revised) graphical view. *WordPress, e-Literate, 10*.

Hood, N., Littlejohn, A., & Milligan, C. (2015). Context counts: How learners' contexts influence learning in a MOOC. *Computers & Education, 91*, 83-91.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 245-251). IEEE.

Jiang, S., Williams, A., Schenke, K., Warschauer, M., & O'dowd, D. (2014, July). Predicting MOOC performance with week 1 behavior. In *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 273-275). International Educational Data Mining Society.

Kesim, M., & Altınpulluk, H. (2015). A theoretical analysis of MOOCs types from a perspective of learning theories. Procedia-Social and Behavioral Sciences, 186, 15-19.

Kizilcec, R. F., & Schneider, E. (2015). Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer-Human Interaction, 22*(2), 6.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60-65).

Lajoie, S. P., & Azevedo, R. (2006). Teaching and learning in technology-rich environments. In *Handbook of Educational Psychology*. Abingdon, UK: Routledge.

Laveti, R. N., Kuppili, S., Ch, J., Pal, S. N., & Babu, N. S. C. (2017, August). Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing. In *Proceedings of the 5th National Conference on E-Learning & E-Learning Technologies* (pp. 1-6). IEEE.

Lee, D., Watson, S. L., & Watson, W. R. (2019). Systematic literature review on self-regulated learning in massive open online courses. *Australasian Journal of Educational Technology, 35*(1).

Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology, 44*(2), 328-337.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 431-439).

McLachlan, G., Do, K. A., & Ambroise, C. (2005). Analyzing microarray gene expression data (Vol. 422). Hoboken, NJ: John Wiley & Sons.

Mezaour, A. D. (2005). Filtering web documents for a thematic warehouse case study: eDot a food risk data warehouse (extended). In *Intelligent Information Processing and Web Mining* (pp. 269-278). Springer, Berlin, Heidelberg.

Michinov, N., Brunot, S., Le Bohec, O., Juhel, J., & Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education, 56*(1), 243-252.

Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017, March). A neural network approach for students' performance prediction. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference* (pp. 598-599). ACM.

Panadero, E., Klug, J., & Järvelä, S. (2016). Third wave of measurement in the self-regulated learning field: when measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research, 60*(6), 723-735.

Papamitsiou, Z., Economides, A. A., Pappas, I. O., & Giannakos, M. N. (2018, March). Explaining learning performance using response-time, self-regulation and satisfaction from content: an fsQCA approach. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 181-190). ACM.

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining, 7*(2), 1-19.

Rhode, J. (2009). Interaction equivalency in self-paced online learning environments: An exploration of learner preferences. *The international review of research in open and distributed learning*, *10*(1).

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological bulletin, 138*(2), 353.

Ruipérez-Valiente, J. A., Cobos, R., Muñoz-Merino, P. J., Andujar, Á., & Delgado Kloos, C. (2017, May). Early prediction and variable importance of certificate accomplishment in a MOOC. In *Proceedings of the 5th European Conference on Massive Open Online Courses* (pp. 263-272). Springer, Cham.

Sun, Z., Xie, K., & Anderman, L. H. (2018). The role of self-regulated learning in students' success in flipped undergraduate math courses. *The Internet and Higher Education, 36*, 41-53.

Tempelaar, D., Rienties, B., & Nguyen, Q. (2018, March). Investigating learning strategies in a dispositional learning analytics context: the case of worked examples. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 201-205). ACM.

Terras, M. M., & Ramsay, J. (2015). Massive open online courses (MOOCs): Insights and challenges from a psychological perspective. *British Journal of Educational Technology, 46*(3), 472-487.

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*(8), 1131-1150.

Willging, P. A., & Johnson, S. D. (2009). Factors that influence students' decision to dropout of online courses. *Journal of Asynchronous Learning Networks, 13*(3), 115-127.

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G. J., & Paas, F. (2019). Supporting Self-Regulated Learning in Online Learning Environments and MOOCs: A Systematic Review. *International Journal of Human–Computer Interaction, 35*(4-5), 356-373.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior, 58*, 119-129.

Xing, W., & Du, D. (2018). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research*.

Ye, C., Kinnebrew, J. S., Biswas, G., Evans, B. J., Fisher, D. H., Narasimham, G., & Brady, K. A. (2015, March). Behavior prediction in MOOCs using higher granularity temporal information. In *Proceedings of the 2nd (2015) ACM Conference on Learning@ Scale* (pp. 335-338). ACM.

Zheng, S., Rosson, M. B., Shih, P. C., & Carroll, J. M. (2015, February). Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1882-1895). ACM.