IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 29, 2019, accepted December 22, 2019, date of publication January 1, 2020, date of current version January 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963503

Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics

PEDRO MANUEL MORENO-MARCOS[1], TING-CHUEN PONG[2], PEDRO J. MUÑOZ-MERINO[1], (Senior Member, IEEE), AND CARLOS DELGADO KLOOS[1], (Senior Member, IEEE)

[1]Department of Telematics Engineering, Universidad Carlos III de Madrid, 28911 Leganés, Spain
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

Corresponding author: Pedro Manuel Moreno-Marcos (pemoreno@it.uc3m.es)

This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R, in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307, a project which is co-funded by the European Structural Funds (FSE and FEDER), in part by the Ministerio de Ciencia, Innovación y Universidades under Grant FPU016/00526 and Grant EST18/00554, in part by the Hong Kong RGC's Theme-Based Research Scheme under Grant T44-707/16-N, and in part by the Innovation and Technology Fund under Grant ITS/388/17FP.

ABSTRACT The advancement of learning analytics has enabled the development of predictive models to forecast learners' behaviors and outcomes (e.g., performance). However, many of these models are only applicable to specific learning environments and it is usually difficult to know which factors influence prediction results, including the predictor variables as well as the type of prediction outcome. Knowing these factors would be relevant to generalize to other contexts, compare approaches, improve the predictive models and enhance the possible interventions. In this direction, this work aims to analyze how several factors can make an influence on the prediction of students' performance. These factors include the effect of previous grades, forum variables, variables related to exercises, clickstream data, course duration, type of assignments, data collection procedure, question format in an exam, and the prediction outcome (considering intermediate assignment grades, including the final exam, and the final grade). Results show that variables related to exercises are the best predictors, unlike variables about forum, which are useless. Clickstream data can be acceptable predictors when exercises are not available, but they do not add prediction power if variables related to exercises are present. Predictive power was also better for concept-oriented assignments and best models usually contained only the last interactions. In addition, results showed that multiple-choice questions were easier to predict than coding questions, and the final exam grade (actual knowledge at a specific moment) was harder to predict than the final grade (average knowledge in the long term), based on different assignments during the course.

INDEX TERMS Prediction, MOOCs, learning analytics, learners' grades, edX.

I. INTRODUCTION

In the past years, many researchers have developed predictive models using MOOC (Massive Open Online Course) data as MOOCs generate big amount of learners' interactions. The main focus has been the detection of academic performance and dropouts, which are often related to the notions of success/failure in the course [1]. The reason is that academic failure in MOOCs is very high and completion rates hardly ever exceed 25% for highly committed learners [2]. Therefore, it is believed that predictions can be used to detect learners at-risk of dropout so as to conduct interventions that could lead to better completion rates.

While there are not many cases of interventions being put in practice in MOOCs, some researchers have started proposing ways to provide feedback and personalization, such as Xing and Du [3], who proposed providing individual drop out probabilities so that instructors could focus on at-risk students and send them e-mail messages (automatically, as it is not feasible to send individual e-mails when there are

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

thousands of learners) to engage them. For such cases, it will be important to provide information to the different stakeholders so managers, instructors and learners are aware of the situation, and they can reflect on what they can do to improve. For example, aggregated data about learners at risk of dropout could be provided to instructors as there are thousands of learners in a MOOC. With this information, they can decide if they want to make any intervention at global scale (for all those learners at risk and/or to groups which share certain behaviors). Moreover, some visualizations could be given to students with the variables that justify that they may be at risk of dropout (e.g., low activity in videos or exercises).

Despite this conceptualization for the use of predictions, one possible issue is that predictive models can be highly context-dependent [4] and researchers provide results that may be only applicable to specific learning environments. A related open research question is to further explore the common factors that influence predictions results so as to be considered in future models to improve the predictive power. If we know the factors that influence prediction and we also know the values of these specific factors for a context, then we can forecast the prediction results and how they might change from contexts.

Among those factors, the variables used to develop the models and the way they are collected (e.g., if variables are collected from the beginning of the course in a cumulative way or if they are collected in a specific period) can be very relevant. As suggested by Moreno-Marcos *et al.* [5], selection of data can be even sometimes more important than the algorithms because variables need to capture appropriate information in relation to the variable to be predicted. Because of that, it is worth analyzing the predictive power of different sets of features. In a MOOC, it can be relevant to analyze to what extent activity in forum can be important as there are thousands of forum posts and learners provide information about different aspects, such as social activity, sentiments and their skills, with the contents of their messages [6]. In addition, other variables should be analyzed as important possible factors such as previous performance and in general, any sets of available variables, as they can add new information.

However, it is important to note that there can also be limitations with data, both for the availability and access to them [7]. This fact can also limit the combination of multiple sources of data [8] and it also raises the question about which variables are enough to predict. If some variables are missing but the predictive power is very high with those present, the effect of the limitations may be alleviated. In contrast, if the best predictors are missing, models could achieve poorer results. In MOOC platforms, it is typical to receive some data containing information such as the results of the activities and more detailed data including all the clicks and events learners perform in the platform (called clickstream). For example, in the clickstream, there is information about every click to play a video, pause a video, move forward, submit an exercise, etc. As this information allows retrieving many additional variables to those obtained with simple results of the activities, it would be relevant to measure their relevance in terms of predictive power.

Another important factor in predictions is the course setup and pedagogy, and aspects such as the course duration, the format of the assignments and the type of questions in the exam may also affect the results. For example, the number of attempts and the persistence [9] (i.e., students' ability to keep on working on the tasks despite the difficulties) can be affected depending on the type of activities. While numerical input questions and open-ended questions (e.g., a programming exercise) can have an unlimited number of attempts, a student automatically knows the answer when the answer is incorrect in a true/false question, and this can affect predictive models.

Furthermore, the prediction outcome that we try to achieve can affect predictions. The prediction outcome we focus in this paper is learners' performance but this can be measured in different ways. Many MOOCs are only evaluated using quizzes and assignments (some of them graded using peer-review) that are usually done during different days or even weeks throughout the course, but others also incorporate a final exam at the end of the course to measure students' level of knowledge. As the tasks are different, the predictive power can vary, and it is relevant to analyze the differences of the predictive power in the different tasks, including the intermediate assignments of the course, the final exam and the final grade. The analysis of the differences between the final exam grades and the final grades based on different activities during the course are also particularly relevant to understand whether or not there are differences between the level of mastery learners show during the course and in a specific moment at the end. In addition, few previous research articles have focused on relating predicting variables, as concluded in a recent review about prediction in MOOCs [10], and it would be relevant to analyze the relationship between the predictions of the final grade the final exam grade, and which of them can be better predicted with available data.

In this context, the aim of the paper is to analyze the influence of different factors in prediction of students' performance, considering intermediate assignment grades (including the final exam) and the final grade. In this direction, the article has been organized into four research questions to guide the analysis. These questions are as follows:

- RQ1) Which factors, among previous grades, forum variables, course duration, type of assignments and data collection procedures, influence grades predictions and to what extent?
- RQ2) How does the presence/absence of clickstream data and interactions with exercises affect prediction results?
- RQ3) How does the question format of the final exam (close-ended and open-ended questions) affect prediction results?

RQ4) Is the predictive power greater for predicting the final exam grade rather than the final grade of the course?

In order to address the abovementioned questions, an analytical method is used in which two MOOCs are compared in order to analyze the influence of the factors in several contexts. The novelty of this paper is because of the following aspects. First, we analyze the effect of how models can be limited when clickstream data is not available. Second, we use a context with final exam in the MOOC and analyze the differences of the predictive power depending on the type of questions. Third, we analyze the relationship between prediction outcomes, as research articles usually focus on predicting separate outcomes, but do not discuss or compare the predictive power between those outcomes. Fourth, we combine the analysis of several factors affecting predictions (use of previous grades, forum data, course duration, type of assignments, and data collection procedure). The analysis of most of these aspects has already been explored in separated works with separated contexts, but we combine all of these aspects in two different contexts to gather further conclusions about how these aspects influence predictions.

The structure of the paper is as follows. Section II presents a background of what have been researched about prediction in MOOCs, and particularly in prediction of grades. Section III presents the methodology used in the study, including the context, data collection, variables, analytical methods and measures. Results of the analysis are provided in Section IV, and discussion is given in Section V. Finally, the main conclusions are highlighted in Section VI.

II. RELATED WORK

In education, researchers have tried to understand and model learners' behaviors, such as retention [11], for years. However, with the introduction of online educational platforms, it is possible to collect more data to better understand what it is happening in the course. For example, [12] found that daily views was a good indicator of student performance and off-campus activity had higher relationship with performance than in-campus activity. With these data, predictive models have been developed in different contexts. One typical context is the courses in higher education institutions. In this case, it is possible to carry out predictions in the whole academic program (typically dropout [13]–[15]) or in specific courses [16], [17]. In the latter case, it is possible to combine data from several sources. For example, Christensen et al. [16], [17] predicted exam outcomes using both self-reported measures and study activities, such as variables obtained in an online platform (Khan Academy in that case) or peer-review activities. Moreover, many researchers have used academic data, such as the CGPA [18]. However, the latter variables cannot be available if we move to fully online contexts, such as MOOCs. MOOCs are different from face-to-face higher education contexts because of the vast number of students, the methodology, and in most cases because of the format, which usually combines videos, exercises, and forum for communication. In this paper, we will focus on the analysis of MOOCs, and from now on, we will focus on what has been specifically researched about prediction in MOOCs.

In MOOCs, there is an increasing interest for research in the last years [10]. For these research works, one of the initial steps is to define the variable to be predicted (i.e., prediction outcome). There are many possible variables to predict, but the prominent ones in the literature are related to learners' behaviors and student performance (e.g., grades, dropout, etc.), although there can be contributions focused on other issues, such as classifying forum posts according to their relevance, [19], etc. Among the first category (learners' behaviors), Bote-Lorenzo and Gómez-Sánchez [20], for example, developed models to predict the decrease of engagement in videos, exercises, and assignments; and Chen *et al.* [21] predicted the personality of learners.

However, most of the contributions have focused on the second category (learners' performance). In this category, dropout prediction has a special relevance because of its high dropout rates. Despite dropouts are not necessarily a problem since many students access the MOOC just because they want to explore some contents [22], there is a high interest on it. In this field, for example, Wu *et al.* [23] developed predictive models using deep learning and neural networks and features obtained through Convolutional Neural Networks (CNN) from raw data. They showed that their method outperformed other traditional machine learning algorithms, such as Random Forest, Support Vector Machines, etc. Particularly, their AUC (Area Under the Curve), which is a common metric to evaluate predictive models, was above 0.9 (which is excellent, according to [10], which also indicates that good AUCs are above 0.8). In addition, Kloft *et al.* [24] predicted dropout over weeks and analyzed how the predictive power improved over time, which suggests there is a trade-off between anticipation (it is desirable to predict as early as possible) and predictive power.

Despite the importance of detecting early dropouts, there are also students who do not drop out the course but they fail. In order to also detect this situation to make interventions, there are models which focus on forecasting learning outcomes, such as grades. In this case, some authors have focused on predicting categorical variables, such as success (i.e., pass/fail) [25], certification at the end of the course [26], and alphabetical grades (e.g., A to D) [27]. In contrast, other works have predicted continuous variables about grades. For example, Ding *et al.* [28] predicted the grade of each chapter of the MOOC using features related to videos and navigation in the MOOC, and Elbadrawy *et al.* [29] the grade of homework assignments using data from previous attempted homework assignments. Pérez-Lemonche, Martínez-Muñoz, and Pulido-Cañabate *et al.* [30], in contrast, focused only on predicting the final grade of the course.

However, there are few contributions focused on predicting grades of the final exam of the MOOC, which is very important as the final exam gives the level of the students after working on the course. Among those, Ren *et al.* [31] used multi-regression models to predict the final exam in a MOOC about Computer Networking, and Pardos *et al.* [32] used Bayesian Knowledge Tracing to forecast students' answers in multiple tests, including the final exam. In addition, research articles often predict one variable or they predict several ones separately, but there is not an analysis of how prediction outcomes are related, which is relevant to understand how predictive models and predictive power differ when changing the prediction outcome [10]. In this paper, we will work in this field (in RQ4), and we will analyze the relationship between the final exam grade and the final grade. This can be useful to understand whether or not students' final grades match with the knowledge they show in the final exam, and it will contribute to the analysis of the relationship between prediction outcomes, which is a general aspect that needs further analysis in the literature.

After addressing the variables to be predicted, another important aspect are the variables used to create the predictive models (i.e., features or predictors). Researchers often transform the raw events into high-level variables to introduce them in the models. Despite there can be many possible variables and new variables can be also explored (e.g., Maldonado-Mahauad *et al.* [33] started exploring the effect of self-regulated learning features), it is frequent to see (1) variables related to the activity in the platform, (2) variables related to forum activity, (3) demographic variables, and variables related to interactions with (4) exercises and (5) videos.

In the first category (activity), Alamri *et al.* [34], for example, achieved promising accuracies (accuracy metric was between 0.82 and 0.92) when predicting dropout using only two features related to the activity, the time spent to complete each content (articles, images, videos) and the total number of accesses to the contents. Among variables related to forum, some examples, used by Klüsener and Fortenbacher [35], are the number of messages, number of words in the messages, and number of ratings emitted/received. However, they are opposite findings about their effectiveness in the predictive models of performance as forum variables can be useful in some contexts (e.g., [36]) but not in others (e.g., [5]). For example, [36] found that variables related to the quality/content of the posts, such as the length of posts, were useful predictors for dropout, while forum variables were not good predictors at all in the research by Moreno-Marcos *et al.* [5] when predicting assignment grades (even using the same variables). This raises the need to further analyze this kind of variables in more contexts. Similarly, demographics variables have also been used in predictive models, although they may show low predictive power respect to variables obtained from learners' interactions, as suggested by Brooks *et al.* [37]. In their case, they found low predictive power of variables such as gender, age, race, language capabilities, geolocation information, whether a learner was paying for the course or not, and signup date when they predicted which learners were going to pass.

In contrast, the variables that have been usually proved to be very effective are those related to exercises. As an example where this kind of variables stand out, Ruipérez-Valiente *et al.* [38] predicted certificate earners in a MOOC and found that the strongest predictor was the progress in problems (i.e., average grade in completed assignments, which could be multiple choice test of peer review activities in that MOOC). In addition, Ren *et al.* [31] predicted assignment scores using different variables and found that the number of quizzes the student took before the assignment was the strongest predictor. Finally, variables related to videos are also relevant in prediction in MOOCs. For example, Yang *et al.* [39] combined assessment grades with variables about videos interactions (using clickstream data), such as the percentage of video that the student played, the average playback rate, the number of rewinds, etc. They also showed that the combination of different features could enhance the predictive power.

Apart from the features, there are other aspects that are also important when developing predictive models. One of them is the source where data comes from. Typically, data are gathered from an online platform, such as edX or Coursera, although other sources (e.g., surveys or an external tool) can be used [10]. For example, Pérez-Sanagustín *et al.* [40] achieved accurate predictions using data outside the MOOC (from an external tool to support self-regulated learning and capture data of activity outside the MOOC, such as visited websites). Moreover, it is important to differentiate between data about student information and progress, and clickstream data (i.e., data containing all the events student do in the platform, such as play, pause, stop a video, submit the solution of an exercise, etc.) since the latter is not always available for the analysis (depending on the agreements with the platform). Some authors, such as Brinton and Chiang [41] already used clickstream data to forecast if users were going to have their answers Correct at their First Attempt (CFA) in a Coursera MOOC about computer networks. They found that the use of features about video-watching behaviors were useful to enhance predictive models respect to the performance obtained with models based on quiz results (data about progress). However, clickstream data is not always available, and, in general, it is not always possible to combine different features because some of them may be missing in some contexts (e.g., a MOOC with just a final exam at the end would not allow collecting information about partial performance). Because of that, it is important to understand which sets of features are the best predictors and which variables are enough to predict with appropriate results (e.g., if clickstream data is necessary to achieve accurate predictions, which is addressed in RQ2).

In relation to the last aspect, the course context is very important (which can limit other aspects such as features, as seen before). In order to delve into different contexts, some authors have analyzed different MOOCs to understand the

differences of the predictive models in different contexts. For example, Qiu *et al.* [42] developed predictive models in 11 MOOCs (five sciences and six non-sciences MOOCs) to predict certificate earners and assignment grades using demographic variables and variables related to forum activity and learning behaviors. They achieved strong predictions in both sciences and non-sciences courses. Given that models may not work when changing the context, others have evaluated the generalizability of the models. This is very important to guarantee a long-term sustainability of the models [43] (i.e., models need to be valid not only for the course used to train the models, but for other courses, as generating models to only be used once would limit their use and their regularly adoption in the institutions in the long term). The topic of generalizability is of real interest nowadays and a recent special issue about early prediction of student performance highlights it as one of the main current research topics [44]. Among research in this field, Boyer and Veeramacheni [45] considered three editions of a MOOC (Spring 2012, Fall 2012, and Spring 2013) and tried to develop predictive models to be used in consecutive editions (e.g., train using data from Spring 2012 and make predictions on learners in the Fall 2012 or Spring 2012 edition). However, they found that the predictive power got worse, which suggests that it is difficult to find generalizable models. In addition, Kizilcec and Halawa [46] trained models using 20 MOOCs and they suggested that using data from multiple MOOCs can improve the generalizability.

However, apart from making models generalizable to other MOOCs, it is important to analyze to what extent the factors that influence those models are the same in different contexts. In this line, there are few contributions. In one of them, Gardner *et al.* [47] evaluated the results of a previous work of dropout prediction in other MOOCs and found that some research results (regarding algorithms and features used to predict) were the same while others indicated just the opposite. In order to get more insight about the factors influencing the predictions, this work will analyze the influence of different factors (mainly in RQ1) and several of them are replicated in two MOOCs as part of the methodology to delve into this issue. For example, one of these factors is about how predictive power differ depending on the type of questions. An early work concluded that open-ended assignments are harder to predict than close-ended ones [5] (typically more concept-oriented) with variables mainly about exercises, but further research is needed in that area, and particularly in the final exam, where the context is different (e.g., different way to ask the question, length of the task, way to evaluate, etc.). This will be addressed in RQ3.

In summary, this paper aims to contribute with the analysis of different factors that can influence predictions of grades. First, factors such as the influence of previous grades, forum variables, course duration, type of assignments, and data collection procedures are addressed. These factors have been linked together because they can be validated with a second MOOC, and they will provide insight about what factors are more relevant when designing predictive models so as to be considered in the implementation. Second, this paper aims to contribute with a deeper analysis of which features are more relevant in the predictions and how the predictive power can be limited when some of them are missing, particularly focusing on multiple edX sources, which can be of interest because of the vast number of MOOCs based on the edX platform. Previous works have focused on which variables have been the best predictors (e.g., [38]), but it is also important to be aware of the effects of data source limitations (i.e., limitations produced when not all the data are available) in the predictive models. This is relevant as not all kind of variables can be always gathered (e.g., there can be a context where interactions with videos cannot be obtained) and models often need to be developed with limited datasets. Third, this paper incorporates a novel analysis focused on the final exam and its different components to delve into the differences on the predictive power depending on the kind of questions. Finally, the paper will innovate with the analysis of the relationship of the final grade and the final exam. This contributes with the analysis of the relationship between prediction outcomes, where more research is needed [10].

III. METHODOLOGY

In this paper, data from two MOOCs are considered. One of the MOOCs was already analyzed in a previous work in terms of prediction [5]. In this work, we take the results from this previous work [5] just for RQ1 and compare them with the results of the new MOOC, which was not considered before. This way, some factors of predictions can be validated in at least two MOOCs. In addition, we add new analyses for the new MOOC from RQ2 onwards. These new analyses are only done with the new MOOC because the methodology of the course and available data does not allow these analyses in the MOOC analyzed our previous work [5], and they add new insight about factors influencing predictive models. This section aims to introduce the context of those courses, the variables and data collection for the analysis, and the analytical methods and measures, including the algorithms and metrics to evaluate the results, and the way to measure the dependent variables of the analyses.

A. COURSE CONTEXT

This study was carried out using data from two MOOCs about Java Programming hosted in edX by two different institutions. The first MOOC is called *Introduction to Programming with Java – Part 1: Starting to Program in Java*, and it was developed by Universidad Carlos III de Madrid (UC3M). This MOOC is the first one of a trilogy of courses for learning Java from scratch. The analysis of this MOOC can be found in a previous article, and this paper will refer to the results presented there [5] and will compare these results with respect to the second MOOC (in RQ1) as way to validate findings. The second MOOC is called *Introduction to Computing with Java*, which was developed by the Hong Kong University of

TABLE 1. Comparison of the characteristics of the two MOOCs.

Characteristic	UC3M	HKUST
Duration	5 weeks	10 weeks
Approach	Concept-oriented	Lab-oriented
Delivery mode	Instructor-paced	Instructor-paced
No. short exercises	135	42
No. summative assignments	7	8
Number of videos	69	122
Enrolled learners	95,555	47,178
Learners watching videos	24,055	19,903
Learners attempting exercises	16,317	11,708
Number of passed learners	1,507	1,299
Passing rate	60%	50%
Time between assignment is opened and deadline	2 weeks (except week 1 where there are 3 weeks)	9 days (excepting project, which is 37 days)

Science and Technology (HKUST) and will be used for the whole analysis.

Despite both MOOCs are on the same topic, the course design and methodology are different, and that could affect the predictive models. A comparison of the characteristics of the MOOCs can be found in Table 1.

Table 1 shows that the duration of the HKUST MOOC is double (10 weeks) than the UC3M MOOC (5 weeks). There are significant overlaps between the topics covered by both MOOCs although some topics are covered at different levels of details. Both MOOCs were instructor-paced, so the contents were released gradually and learners had specific deadlines to complete the assignments, although the pedagogical approaches were different. UC3M MOOC was more focused on the concepts, and it was evaluated through seven assignments, which consisted of five close-ended tests (one each of the five weeks of the course, counting 15% each) and two peer-review programming assignments in week 3 and 5 (counting 10% and 15%, respectively). Moreover, it contained 135 close-ended short formative questions which do not count for the final grade, which needed to be above 60%, the passing rate, to pass (regardless the grade of specific items).

In contrast, HKUST MOOC was more lab-oriented, which assesses mastery of concept as well as skill. In that MOOC, labs consisted on several programming tasks (typically 4-5 exercises for each lab) in which students needed to submit their Java code, which was automatically assessed. The evaluation consisted of these automated graded labs (which were six and the top 5 grades counted 20%,) plus one project (with one automated graded part and another with peer-review, which counted 20%), and a final exam, which counted 40% of the final grade. In this case, students needed to get at least 50% of the points (regardless the grade of specific items) to pass. The six labs were part of weeks 1, 2, 3, 5, 7 and 8, the project was released on week 5 although the deadline was on week 9, and the final exam was administered on the last week (week 10). Additionally, the MOOC contained 42 short exercises (both conceptual, e.g., multiple-choice questions, and about coding, e.g., writing a short piece of code) intended to prepare learners for the assignments. These short exercises were automatically graded and counted towards the remaining 20% of the final grade.

The number of videos is also higher in the HKUST MOOC, which could be expected as the duration is higher. Regarding the population, enrolled users in the UC3M MOOC is double that of HKUST, although the number of learners who actually passed the course and engaged in the activities in the UC3M MOOC was only slightly higher. The demographics about the two MOOCs are also similar, with a median age of 30 years and 70% of learners are university degree holders for both MOOCs. Geographically-wise, figures were also similar. Most students came from USA in both MOOCs (25% in both MOOCs), followed by India (18% in the HKUST MOOC and 16% in the UC3M MOOC). The third country with more learners was Spain in the UC3M MOOC (3%), while United Kingdom in the HKUST MOOC (3%).

B. VARIABLES AND DATA COLLECTION

In order to analyze the data of the MOOCs, data from edX have been collected. Particularly, the following sources from edX have been considered [48].

- *{org}-{course}-{run}-course_structure-{site}-analytics.json*: Contains the structure of the course and all its assessment items.
- *{org}-{course}-{run}-courseware_studentmodule-{site}-analytics.sql*: Contains the state of learners in each course components.
- *{org}-{course}-{run}-{site}.mongo*: Contains the forum messages and the information about the forum interactions.
- *{course_id}-grade_report_{datetime}.csv*: Contains the grades of the learners in each assessed activity.
- *{org}-{site}-events-{date}.log.gz.gpg*: Contains the low-level users generated events when they interact with the platform, i.e., clickstream data.

The first four files are available in both MOOCs and they serve to obtain variables related to the interactions with exercises (second file) and forum (third file) in the different parts of the course (following the structure in the first file). The second file also gives information about which videos have been opened, but not about how much they have been watched. Comprehensive information about interactions with videos and activities, including the events of clicks learners had with the platform (e.g., when user plays a video, pauses it, etc.), i.e., clickstream data, is obtained using the fifth file. However, last file is only available in the HKUST MOOC.

With regard to the methodology for using the data, in order to have comparable results with those obtained in the UC3M MOOC, the sample selection criteria for the HKUST MOOC are the same as in the UC3M MOOC [5], and consists on

TABLE 2. Variables used in the study.

Variables	Description
Forum	
participations	No. messages a learner posted
com_threads	No. threads a learner started
comments	No. answers in threads a user posted
votes_rcv	No. votes a learner received from others
votes_emt	No. votes a learner emitted to others
endorsed	No. messages flagged as valuable/relevant
avg_length	Average number of characters of posts
percentile	Percentile about participations in the forum
sentiment	Average positivity of messages (see [5])
Exercises	
avg_grade	Average grade of short exercises (non-attempted are 0)
attempted	% of exercises attempted
opened	% of exercises opened (although not submitted)
avg_attempts	Average number of attempts in the exercises attempted
per_correct	% of exercises with 100% correct
CFA	% of exercises 100% Correct at First Attempt (CFA)
Videos	
per_open	Percentage of opened videos
per_vtotal	Viewed percentage of total video time
per_compl	Percentage of completed videos
avg_rep	Average number of repetitions per video
avg_pause	Average number of pauses per video
Activity	
streak_acc	Longest consecutive run of accesses to the platform
ndays	No. days the student has accessed to the platform
avg_con	Average no. of consecutive days accessing to the course
per_pc	Percentage of accesses from a PC
per_wk	Percentage of accesses during weekend
Previous grades	Grades from available previous graded assignments

filtering students who did not engage with the course and they did not participate in the forum. This produces a filtered dataset of 2,168 learners in the HKUST MOOC. Moreover, the same variables are collected to ensure the replicability of the experiments in RQ1. Nevertheless, related to RQ2, additional variables are collected from the clickstream (fifth file) to analyze whether they can enhance the predictive models (these variables were already used in another previous work [43]). Table 2 presents the list of independent variables used in the study. In that table, variables in bold are obtained from the clickstream and thus they are only available in the HKUST MOOC.

These variables can be gathered using two modes, which are worth analyzing to discover the best way to collect data so as to capture the most representative variables. These modes are the cumulative and non-cumulative mode. Cumulative mode means that they are collected from the beginning of the course until a specific moment (e.g., end of week X), and non-cumulative means that they are collected within a specific period (e.g., from the release of week X to the deadlines of that week). In the analysis, we will refer to the general term ''data collection procedure'' to refer the mode use to collect the data.

C. ANALYTICAL METHODS AND MEASURES

In order to develop predictive models, it is necessary to define the predictors. However, it is also important to define the variables to be predicted, i.e. the prediction outcomes. In this case, the prediction outcomes are the grades of each of the graded assignments of the MOOC (including the final exam). There are seven graded assignments in the UC3M MOOC (five close-ended tests and two programming assignments) and nine in the HKUST MOOC (six labs, the automatic graded component of the project, the peer-review component of the project, and the final exam). These grades are measured in scale 0-1, as they appear in the grade report. Moreover, the learning outcome (pass/fail) is also considered in the analysis when analyzing the prediction of the final grade and the final exam. For this case, a binary variable is used, whose value is 1 if the grade is above the passing rate, and 0 otherwise.

With these definitions of the variables, predictive models are generated using the *sklearn*[1] library of Python in the Anaconda platform, and four well-known predictive algorithms: (1) Regression (RG), (2) Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel and epsilon = 0.01, (3) Decision Trees (DT), and (4) Random Forest (RF) with 500 estimators. With these algorithms, results are retrieved using 10-fold cross validation. The parameters of the algorithms not mentioned before are the default values in the *sklearn* implementation.

In order to measure the performance of these algorithms, RMSE (Root Mean Square Error) has been used when predicting scores, as suggested by Pelánek [49], who preferred this metric respect to others, such as MAE (Mean Absolute Error), which are biased towards the majority result. Moreover, AUC is used when predicting the learning outcome, as it is generally appropriate for student behavior classification problems [49], and avoids some problems in imbalanced datasets [50].

Despite RMSE and AUCs are recommended for this context and they will be used to report results in Section IV, an Appendix is included with the results using other metrics. The reason for this appendix is that many articles often use other metrics (although these metrics may not be the best) and this appendix will allow researchers comparing their results when other common metrics in the literature are used. Particularly, two metrics are included: MAE for grades' predictions and F-score for pass/fail predictions. In the Appendix, it can be seen that the main conclusions in the educational field are the same, but it is relevant to report other metrics to better compare research results.

IV. RESULTS

This section is divided into four subsections according to the four research questions.

A. RQ1: INFLUENCE OF PREVIOUS GRADES, COURSE DURATION, TYPE OF ASSIGNMENTS, FORUM VARIABLES AND DATA COLLECTION PROCEDURE

The first part of the analysis aims to analyze the influence of different factors that can affect the predictive power when predicting learners' performance in assignment grades

[1] https://scikit-learn.org/stable/documentation.html

(including the final exam) of a MOOC. In particular, the prediction factors that are addressed in this section are as follows:

a) The influence of previous performance on graded assignments on the prediction of future assignment grades.
b) The influence of course duration on the prediction of assignment grades.
c) The influence of the type of assignments (concept-oriented vs. lab-oriented) in the predictive power of assignment grades.
d) The influence of forum variables on the predictive models to predict grades of assignment grades.
e) The influence of the data collection procedure (cumulative vs. non-cumulative, as explained in Section 3.B) in the predictive power.

In order to analyze these factors, six models have been created using different sets of variables for the two MOOCs. These models were also used in a previous work [5], so it is possible to compare the results because the same conditions are applied. For Models A-E, variables are collected in non-cumulative mode (with variables collected within the week where the assignment is due). In contrast, Model F uses cumulative mode (with variables collected from the beginning of the course to the deadline of each assignment). The purpose of Model F is to compare the data collection procedure and it is used with the same variables as Model A, following the methodology used in a previous work [5]. A summary of these models is as follows:

- Model A: Variables related to problems and *per_open*
- Model B: Model A plus previous grades of graded assignments.
- Model C: Variables related to forum.
- Model D: Model A plus variables related to forum.
- Model E: Model B plus variables related to forum.
- Model F: Variables of Model A in cumulative mode.

With each model, grades of the different graded assignments have been predicted. For the case of the UC3M MOOC, grades are predicted for the close-ended tests (T_i), the programming assignments (P_i) and the final grade (FG), where i indicates the week where each assignment is placed. Results for this MOOC are taken from a previous article [5] and can be found in Table 3. These are the only results taken from another article. The rest of them are new. For the case of the HKUST MOOC, grades are predicted for the six labs (L_i), the project submission (PS) with automatic grading, the project submission with peer-review (PR) and the final exam (FE). The week where assignments are due is indicated as W_i. These results are in Table 4. Note that FG is not predicted yet in the HKUST MOOC because variables related to exercises cannot be used as all of them count towards FG (20%). This is different in the UC3M MOOC where there are 135 short formative exercises which do not count for the FG at all.

A first look at Tables 3 and 4 shows that the best values in both tables are in Models B and E, which are the models that include the previous grades of graded assignments. This finding suggests that previous performance (and particularly in graded tasks) is among the strongest predictors of future performance, as also shown in other contributions (e.g., [38] found grades in assignments was the best predictor for dropout). Moreover, another common finding in both MOOCs is that the predictive power tends to be better in the last assignments. This can probably be because as the course

TABLE 3. Results of the predictive models in the UC3M MOOC.

Method		T1	T2	T3	T4	T5	P3	P5	FG
Mod. A	RG	.27	.22	.20	.18	.16	.25	.20	.14
	RF	.26	.21	.20	.19	.16	.25	.21	.14
	SVM	.27	.22	.21	.19	.17	.25	.21	.15
	DT	.34	.28	.26	.22	.18	.31	.27	.16
Mod. B	RG	.27	.21	**.18**	**.15**	**.13**	**.24**	**.19**	-
	RF	.26	**.20**	**.18**	**.15**	**.13**	**.24**	**.19**	-
	SVM	.27	**.20**	**.18**	**.15**	**.13**	**.24**	.20	-
	DT	.34	.26	.23	.20	.17	.32	.26	-
Mod. C	RG	.41	.36	.33	.31	.27	.34	.24	.26
	RF	.42	.37	.33	.31	.27	.34	.25	.25
	SVM	.46	.40	.35	.33	.28	.34	.26	.28
	DT	.46	.40	.35	.33	.30	.36	.28	.27
Mod. D	RG	.26	.22	.20	.18	.16	.25	.20	.14
	RF	**.25**	.21	.20	.19	.16	.25	.20	.14
	SVM	.26	.22	.21	.19	.17	.26	.21	.15
	DT	.34	.28	.26	.23	.19	.32	.28	.17
Mod. E	RG	.26	.21	**.18**	**.15**	**.13**	**.24**	**.19**	-
	RF	**.25**	**.20**	**.18**	**.15**	**.13**	**.24**	**.19**	-
	SVM	.26	**.20**	**.18**	**.15**	.14	.25	.20	-
	DT	.34	.26	.23	.20	.17	.32	.26	-
Mod. F	RG	.27	.25	.24	.23	.21	.27	.22	.15
	RF	.26	.23	.22	.21	.18	.26	.22	**.13**
	SVM	.27	.23	.22	.22	.19	.27	.22	.14
	DT	.34	.31	.30	.29	.25	.36	.30	.17

Note: Best RMSE values per column are in bold.

TABLE 4. Results of the predictive models in the HKUST MOOC.

Method		L1 W1	L2 W2	L3 W3	L4 W5	L5 W7	L6 W8	PS W9	PR W9	FE W_{10}
Mod. A	RG	.33	.33	.28	.30	.26	.28	.26	.25	.17
	RF	.34	.36	.30	.33	.28	.30	.27	.26	.18
	SVM	.36	.36	.29	.33	.27	.34	.29	.26	.19
	DT	.38	.41	.33	.36	.32	.31	.28	.27	.18
Mod. B	RG	.33	**.31**	**.26**	.27	**.24**	**.26**	**.25**	**.24**	**.16**
	RF	.34	.33	.27	.27	**.24**	.27	.26	.25	.17
	SVM	.36	.35	.27	.27	**.24**	.28	.27	.25	.17
	DT	.38	.38	.31	.30	.29	.30	.30	.28	.22
Mod. C	RG	.40	.45	.47	.45	.43	.38	.33	.37	.38
	RF	.42	.47	.50	.46	.43	.39	.33	.37	.38
	SVM	.45	.56	.57	.53	.48	.42	.35	.41	.45
	DT	.47	.53	.56	.50	.46	.42	.35	.38	.39
Mod. D	RG	**.32**	.33	.28	.30	.26	.29	.26	**.24**	.17
	RF	**.32**	.34	.29	.33	.28	.30	.27	.26	.18
	SVM	.33	.36	.29	.34	.27	.33	.29	.26	.19
	DT	.39	.45	.38	.39	.33	.34	.31	.30	.19
Mod. E	RG	**.32**	**.31**	**.26**	**.26**	**.24**	**.26**	**.25**	**.24**	**.16**
	RF	**.32**	.32	.27	.27	**.24**	.27	.26	.25	.17
	SVM	.33	.34	.27	.27	**.24**	.29	.27	.25	.17
	DT	.39	.40	.35	.33	.30	.31	.32	.30	.22
Mod. F	RG	.33	.34	.32	.30	.30	.31	.30	.27	.23
	RF	.34	.36	.31	.29	.26	.28	.28	.25	.19
	SVM	.36	.35	.31	.29	.26	.28	.29	.26	.20
	DT	.38	.46	.42	.40	.36	.38	.38	.35	.26

Note: Best RMSE values per column are in bold.

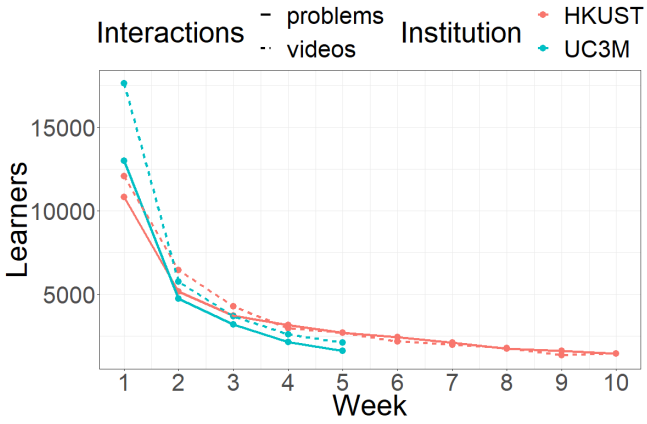

FIGURE 1. Evolution of the number of learners engaging with videos and exercises.

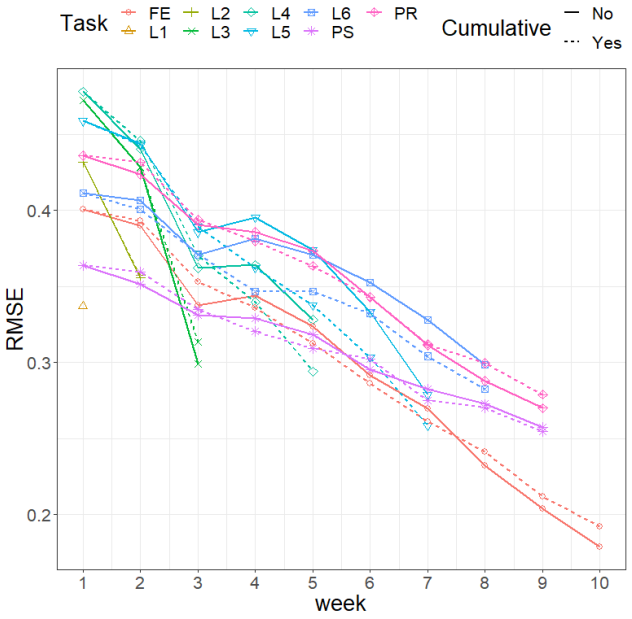

FIGURE 2. Prediction of grades using data from previous topics in the HKUST MOOC.

evolves, more data are available and less students interact with the course, which makes it is easier to separate between those who are engaged and those who are not. However, at the beginning of the course, there can be learners who "sample" some activities to explore the course but they are not committed to do it (known as sampling learners by Maldonado-Mahauad *et al.* [51]).

For these MOOCs, an important difference is the duration, which is double in the HKUST MOOC. However, the engagement (measured with the number of learners interacting with the MOOC components) follows a similar exponential pattern for both videos and exercises (Fig. 1), with a prominent drop in the first 2-3 weeks. This entails that although the number of committed learners drops every week and that may affect the longer course, there are not many differences due to the course duration since the most critical weeks are the first ones in both MOOCs. This also supports the fact that the improvement of the predictive power is higher in the first few weeks than in consecutive weeks.

Another significant difference is that the predictive power is generally slightly worse in the HKUST MOOC than in the UC3M MOOC. A possible reason can be related with the type of questions. HKUST MOOC is mostly assessed based on labs (lab-oriented approach). In the UC3M MOOC, the predictive power is significantly higher for graded tests (concept-oriented) than for programming assignments (lab-oriented), and the predictive power of most of HKUST assignments is similar to P3 in UC3M MOOC. The fact that the prediction of FE of HKUST MOOC (which contained 30 multiple-choice concept-oriented questions and lab-oriented coding questions) is significantly higher can also support this finding. The implication is that lab-oriented tasks can be more difficult to predict. A deeper analysis on the differences of the predictive power depending on the type of questions will be provided in RQ3.

With regard to the forum variables, results show that they do not improve significantly the models in neither of the MOOCs for the different algorithms. While learners were highly encouraged to use forum in both MOOCs, the HKUST MOOC also had specific formative activities to be solved in the forum with hundreds of responses (one activity had 826 responses), although they did not contribute to improve the predictive power of forum variables. Moreover, the forum of the HKUST MOOC seemed to work very well with 94% of threads with response and an average time of response of six hours. These facts can also support that forum activity is not directly related to the learning outcomes.

Finally, another difference is related to the use of cumulative/non-cumulative mode. While it is always better to use data of only the current week at UC3M, that does not always happen at the HKUST MOOC. To explore further into this finding, Fig. 2 illustrates the prediction of different assignments using both modes over time (using models A and F with RF, as used in the analysis of the UC3M MOOC [5]). The figure shows that unlike in the UC3M MOOC, it is not clear which mode is better. Non-cumulative mode is better for some tasks but not for others, and there are some tasks (e.g., FE) where the best mode varies along the timeline. A possible reason for this is that the HKUST MOOC has considerably fewer short exercises (there are only 3-4 exercises per week), which can make variables be less representative in non-cumulative mode and achieve worse predictive power.

Moreover, the figure also shows that it is difficult to achieve accurate predictions until the latest stages of the course in the HKUST MOOC. The higher standard deviation between the results of each task for each student in the HKUST MOOC respect to UC3M MOOC (0.29 vs 0.14) can also make grades more difficult to be predicted. This fact is also related to the anticipation. Results show that the predictive power improves week by week and as the predictive power is not very accurate until the last stages, it is not possible to get

very accurate predictions soon. While there is always a trade-off between anticipation and predictive power, these results may suggest that predicting one week ahead could be a good option to avoid losing much predictive power and allowing some anticipation (predictions in the last moment are useful to identify factors but not to make impact on learners).

In summary, both courses have in common that the predictive power improves over the time in a similar way, with previous grades as strong predictors and low predictive power of forum variables. Moreover, both courses share a similar curve of decrease of engagement over time, and not with the proportion of the completion of the courses. However, the predictive power is generally worse in the HKUST MOOC, although the predictive power when predicting the labs is comparable with similar programming tasks in the UC3M MOOC. This suggests that the type of assignment may affect the results, and concept-oriented tasks can be easier to predict, as concluded in a previous work [5]. The prediction mode also presents differences and it is less clear in the HKUST MOOC, perhaps because of the few activities to be considered in the non-cumulative mode and more lab-oriented assessments. This may imply that further analysis can be done in the HKUST MOOC to analyze whether if could be beneficial to include more tasks and explore a better predictive model for lab-oriented skill-based activities.

B. RQ2: EFFECT OF CLICKSTREAM DATA AND VARIABLES RELATED TO EXERCISES ON PREDICTION

The second part of the analysis aims to analyze the effect of the clickstream data (data containing the events generated by learners when they make clicks in the platform). As these data are not always available (e.g., UC3M MOOC), the HKUST MOOC is used for this objective. It is relevant to know their relevance in their predictive models, if their inclusion can improve the predictive power and discover if it is enough with courseware data. In order to analyze this question, two additional models have been defined:

- Model G: Variables obtained from the clickstream (variables in bold in Table 2 plus *per_open*, which is included as it can be also computed with clickstream and it completes the set of variables related to activity and videos), i.e. just using clickstream data.
- Model H: Model E plus variables obtained from the clickstream

On the one hand, Model G has been selected as a base model to know how much predictive power can be achieved with just the information about videos and activity provided by the clickstream. On the other hand, Model H has been chosen as an extension of Model E, which was the more comprehensive model without clickstream and the model which achieved better results. Therefore, Model H would be the most comprehensive model as it contains all the variables.

Note that it is normally possible to collect the same variables collected from the courseware using the clickstream. Therefore, in this case, when we refer to clickstream, we focus on the variables about videos and activity, that can only be retrieved with clickstream data and not with the courseware and other data. However, in this context, it is noteworthy that variables related to exercises cannot be retrieved from the clickstream since many problems are graded with an external grader and the events do not record the grade. This is a limitation, although it is not impeding for the study as they are available in the courseware. Taking the new models into account, results are presented in Table 5.

TABLE 5. Results of the predictive models using the clickstream data in the HKUST MOOC.

Method		L1 W1	L2 W2	L3 W3	L4 W5	L5 W7	L6 W8	PS W9	PR W9	FE W_{10}
Mod. G	RG	.37	.38	.39	.35	.32	.32	.29	.26	.21
	RF	.37	.38	.38	.35	.32	.33	.28	.26	.21
	SVM	.42	.41	.42	.39	.36	.37	.32	.29	.21
	DT	.48	.51	.49	.46	.42	.43	.36	.32	.25
Mod. H	RG	**.32**	**.31**	**.26**	**.26**	**.24**	**.26**	**.25**	**.24**	**.16**
	RF	**.32**	**.31**	**.26**	**.26**	**.24**	**.26**	.26	**.24**	.17
	SVM	.34	.34	.27	.27	**.24**	.29	.27	.26	.17
	DT	.41	.43	.34	.37	.30	.35	.35	.32	.21

Results show that the variables added with the clickstream (related to activity and videos) do not offer a significant increase of predictive power and Model H achieves similar results to Model E, which means that variables obtained from clickstream data do not improve the predictive power and they do not add new insights in the prediction if the variables considered in model E are taken into account. This also implies that the analysis with courseware data (as it was done with the UC3M MOOC) can be enough to achieve predictive power as it captures the best predictors.

In order to delve into the best predictors, the importance of variables has been computed for the RF models with Model H (see details of calculation in the article by Louppe *et al.* [52]). Fig. 3 shows the results of the variables with higher importance (normalized, so the sum of all variables for each task is 1). This figure clearly indicates that variables related to exercises are the best predictors. The average grade is a strong predictor for most of the assignments and particularly in cumulative mode (and this result matches in the UC3M MOOC). Moreover, previous grades are also very strong predictors. Particularly, grade from the previous lab is the best predictor for L4 to L6 and for PR and PS. This highlights the importance of past performance to predict future performance. The fact that the previous grades lose predictive power after the following assignments may mean that the last performance can be the best predictor. The few exercises for each week can suppose a limitation when considering the latest interactions, though. Nevertheless, this result servers to justify that variables related to exercises are crucial to develop the predictive models. If they can be obtained from the clickstream, results can be accurate; otherwise the clickstream without these variables may not be enough, as corroborated by the poorer results of Model G.

However, for the prediction of FG, variables related to exercises cannot be used because all exercises are part of

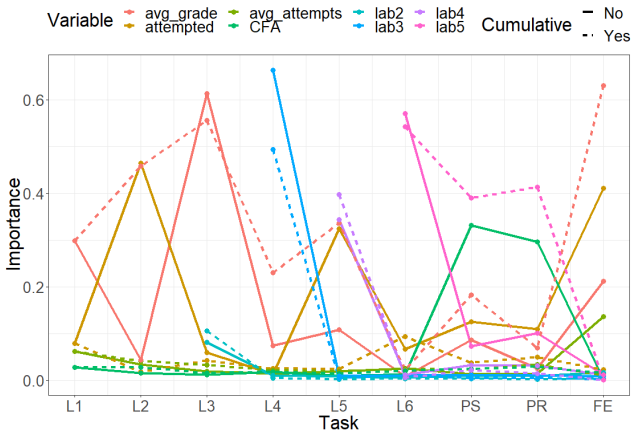

FIGURE 3. Importance of variables for the different tasks of the MOOC.

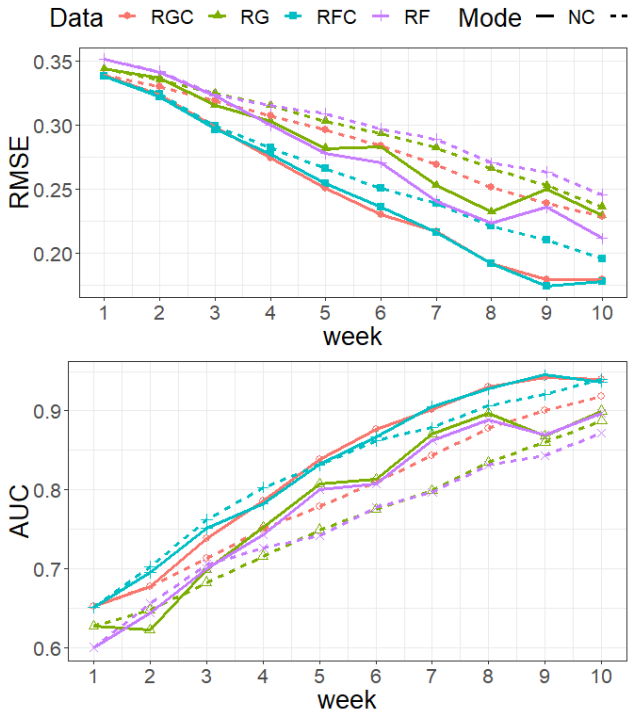

FIGURE 4. Prediction of FG with RG and RF with (RGC and RFC) and without using clickstream data (RG and RF) in cumulative (C) and non-cumulative (NC) mode.

FG. Therefore, models have to be developed using only interactions with videos, forum and the activity in the platform. In this case, a model has been developed using forum variables and *per_open* (videos), which are the only available without the clickstream, and another model also including the clickstream. The evolution of the predictive power for both the prediction of the score and the learning outcome (pass/fail) using the most consistent algorithms in the analysis (RG and RF) are presented in Fig. 4.

Results show that despite clickstream does not considerably enhance the models when variables related to exercises are present (as seen in Table 5), they are relevant when no other information is available. This can be seen (in Fig. 4) with the high difference between models with and without clickstream for the FG where exercises cannot be used. The best RMSE at the end of the course is 0.18, which means that while there are errors in the prediction, it is possible to give a good estimation (in comparison with other works, such as the article by Pérez-Sanagustín *et al.* [40], where the best RMSE was 0.18 at the end of the course) of the range of the grade. Furthermore, prediction of the learning outcome provides better results and it is possible to predict with good results (AUC above 0.8) from week 4 (40% of the course) and excellent (AUC above 0.9) from week 7 (70% of the course). These results are worse than in the UC3M MOOC (good AUC from week 1 and excellent from 20% of course duration), but they can still provide early predictions to be used in live courses.

In summary, variables related to exercises are very relevant for achieving strong predictions. If they cannot be used, the predictive power could still be acceptable in later stages of the course, but the use if preferable. The positive point for edX researchers if that courseware data provides information about exercises, and it is possible to develop strong models with just that information, which is easier to handle than clickstream (in the HKUST MOOC, courseware data was about 600 MB, while clickstream was about 20 GB). Nevertheless, if clickstreams have the information about exercises, their use can also get insight about learners' behaviors while achieving accurate predictions. Moreover, clickstream data could also provide formative feedback to the instructors on improving the instructional design.

C. RQ3: EFFECT OF THE QUESTION FORMAT IN PREDICTION OF THE FINAL EXAM

One particularity of the HKUST MOOC is that it contains a final exam, unlike many other MOOCs. This exam was divided into two main components: 30 multiple-choice questions (MC, 2 points each) and automated-graded coding problems (CD, 10 points each). As previous results suggest that there may be differences because of the type of assignment (e.g., lab grades may be difficult to predict), it is interesting to delve into the differences of the type of questions of the exam, which is the only task which combines close-ended (MC) and open-ended (CD) questions. In order to do that, grades for each component of the exam (MC and CD) have been separated and predictive models have been developed to predict each separate component. Fig. 5 depicts the evolution of the predictive power when predicting the different components of the exam using Model H (which is used because it contains all possible the interactions and it outperforms most of the models) and RG and RF as algorithms (SVM and DT have been removed to make the plot clearer as their predictive power was worse).

Fig. 5 clearly shows that the predictive power of CD is worse than the predictive power of MC and FE. This result matches with previous findings about the difference of prediction between close-ended (as MC) and open-ended

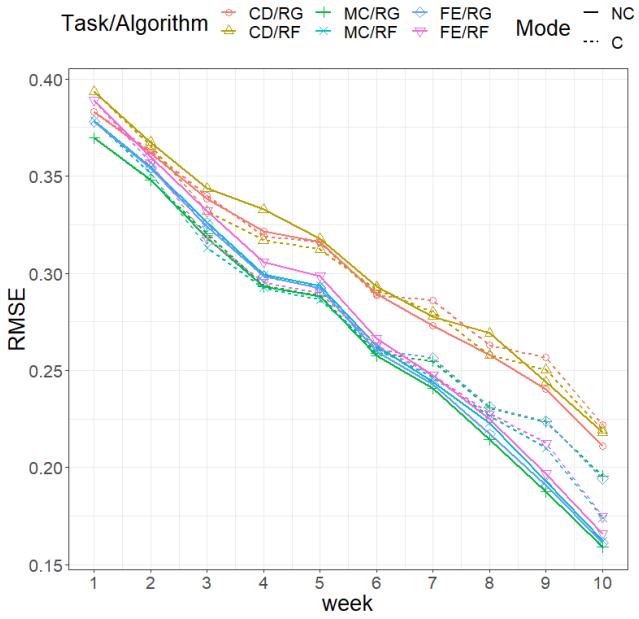

FIGURE 5. Evolution of the predictive power when predicting MC, CD, and FE itself, using cumulative (C) and non-cumulative (NC) mode.

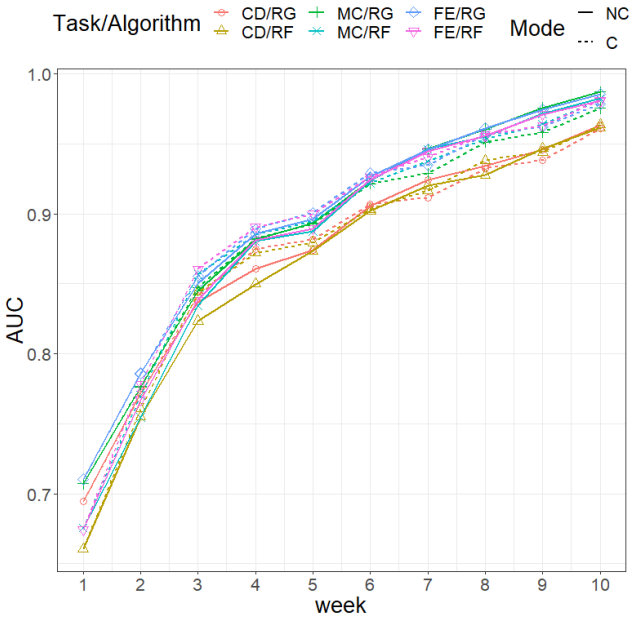

FIGURE 6. Evolution of the predictive power when predicting MC, CD, and FE itself, using cumulative (C) and non-cumulative (NC) mode.

(as CD) questions [5], and the finding we obtained about the predictive power of HKUST assignments, which was generally worse probably because of the type of questions (they were mostly open-ended tasks). In contrast, the predictive power is very similar for the FE and MC. This means that the prediction of FE is not highly affected because of the CD, which are harder to predict, and this is good. Results also show that the non-cumulative mode is better for all the components, which entails that the latest interactions are the best predictors. As mentioned previously, as many learners who are not committed to engage in the course interact at the beginning, initial interactions can introduce noise in the model, and thus the latest interactions are preferable. In terms of the algorithms, RG and RF achieve similar results, as it happened in previous findings, which suggests that strong predictors are linearly related with the outcome. In fact, at the end of the course, the correlation between the variables related to exercises with higher importance in Fig. 3 (*avg_grade*, *attempted*, *avg_attempts*, and *CFA*) is higher than 0.8 for all variables.

Results also show that the best predictive power for both MC and FE is 0.16 at the end of the course and 0.21 for CD. This means that it is possible to give a good estimation of the range of the grade (as happened with FG) for the different components, although predictions of CD are worse. If predictions are only carried out as pass/fail, AUC (see Fig. 6) are above 0.8 for all components from week 3 and above 0.9 from week 6. This entails that while CD can reduce the anticipation and predictive power for the grade, they do not for the learning outcome, which is positive, as the mixture of both kind of tasks can be enriching for making learners master both conceptual and practical skills. In fact, this suppose that parts of FE can be accurate predicted even before half of the course, which is good. Moreover, the fact that accurate results can be achieved from the first weeks means that it is possible to obtain early predictions so as to be used for prevent learners' failure.

D. RQ4: COMPARISON OF THE PREDICTIVE POWER BETWEEN FINAL EXAM GRADE AND FINAL GRADE

In this section, our main objective is to compare the predictive power that we can achieve between FE and FG. In the HKUST MOOC, the FE represented 40% of the FG, so it is expected to be high representative of the FG, but there are some patterns that are worth analyzing. In order to analyze the differences, Fig. 7 represents the relationship between both grades. It can be seen that for passing grades, there is a positive linear relationship between FE and FG (correlation of 0.95 using all points, and 0.58 using only learners who passed the exam). Moreover, there are many learners who did not take the exam. Almost all of them failed the course and probably were dropouts, although there are 12 cases of learners who had reached the minimum 50% required to pass without the FE and they did not take it. Furthermore, there were very few cases in the dataset ($n = 48$) of students who failed the exam, which means that those who took it generally passed, and from those who failed, 52% of them passed the course. Therefore, it can be said that both grades are very highly related and almost all learners who passed the course also passed their FE.

Taking into account that both grades are closely related, it is interesting to analyze the differences of the predictive power of the models to forecast both the FG and the FE depending on the prediction outcome. As one limitation when

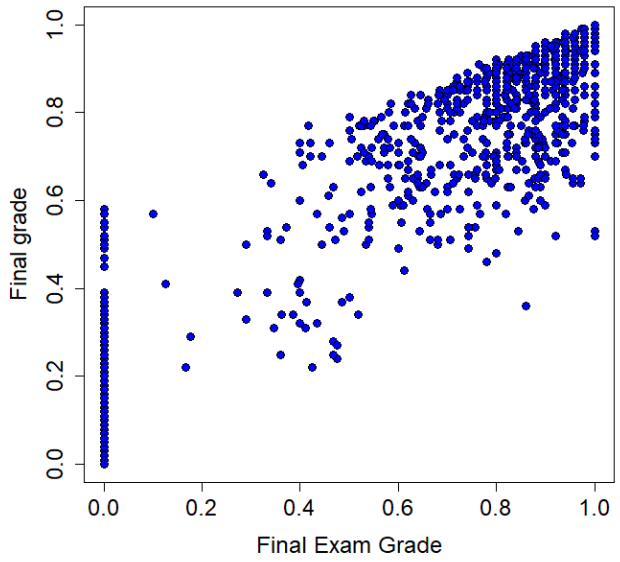

FIGURE 7. Relationship between the FE and FG.

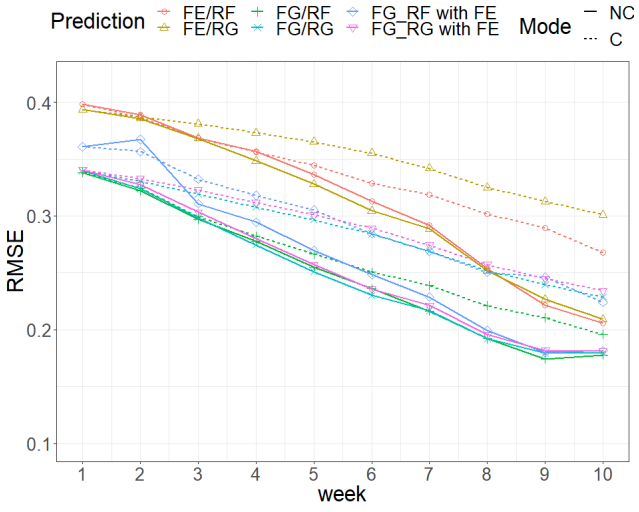

FIGURE 8. Comparisons of predictive models to forecast the FE and FG.

predicting FG is that it cannot be predicted using variables related to exercises because they are part of the FG, models for both prediction outcomes use variables related to activity, videos and forum (as it was done in RQ2 with FG). This way predictions of the FG will comparable to those with the FE as the predictors are the same. Considering this model, the evolution of the predictive power for the FG and FE is presented in Fig. 8. In addition, in order to understand the relationship between the variables, an additional model has been included to predict FG using only the predictions obtained with the model of FE (one predictor variable).

A first observation is that the predictive power of the FG is almost always better than the prediction of the FE, and more time is needed to get accurate predictions of the FE (e.g., RMSE reaches about 0.2 in week 8 for FG and week 10 for FE). This suggests that the average knowledge in the long term is harder to forecast than the actual knowledge at a specific moment. With regard to the data collection procedure, non-cumulative mode seems to be the best for both prediction outcomes, so it should be preferred. Another observation is that the predictive power obtained when predicting FG using only the predictions of the FE is not so different to the model used to predict FG with learners' interactions. While differences are higher in the first stages (particularly with RF), results are very similar in later stages. This entails that due to the high (and linear) relationship between both grades, as shown in Fig. 7, it is possible to predict FG from just the predictions of the FE without losing so much predictive power. This result implicates that prediction outcomes can be sometimes related and further research can be done to analyze the relationship between prediction outcomes, as suggested by [10], which identified this lack of research in their literature review. Nevertheless, the finding about FG is easier to predict than the FE can be relevant and can be considered in the implementation of predictive systems in live MOOCs.

V. DISCUSSION

In this paper, an analysis of students' performance prediction has been carried out analyzing some factors influencing the prediction, some of them related to the predictor variables, others to the prediction outcomes, and others related to the methodology of the analysis or the course. This study took the same methodology as a previous work [5] to conduct the analysis. This means that the same features, algorithms, filtering criteria and predictive models (models A to F), etc. were used in this analysis. However, there were important differences in the analysis of both works that are worth mentioning. First, the analysis of the influence and/or importance of clickstream data (Research Question 2) is new in this contribution. Second, there was not FE in our previous work [5], and therefore the analysis of the FE and its relationship with FG is new (Research Question 4). Third, the typology of closed-ended and open-ended items was different, which made Research Question 3 different from what has done in our previous work [5]. UC3M MOOC had close-ended assignments and peer-review assignments, while in the HKUST MOOC, we only focused in one exam which combined different types of activities. In addition, the nature of open-ended tasks was different. In the UC3M MOOC, open-ended tasks were peer-reviewed, so their evaluation had a subjective factor. In contrast, in the HKUST MOOC, the open-ended tasks we analyzed in Research Question 3 were part of an exam, and they were automatically graded, so grading was more objective. Furthermore, they were small exercises that involved writing single methods whereas the UC3M MOOC open-ended tasks were similar to the labs in the HKUST MOOC (they both involved more complex tasks than coding questions of the HKUST exam, and they usually involved several Java methods). Therefore, their comparison is different in each case, although they share that there were close-ended and open-ended activities in both cases. Finally, the comparison of the results of the two MOOCs

on the same topic but with different teaching methodology (Research Question 1) is new from this paper and relevant to analyze how research results can generalize.

Considering the previous differences, in the two considered MOOCs, it was observed that the best predictors were variables related to exercises and particularly, grades from previous assignments were very strong predictors. This matches with contributions which suggest that previous performance is useful to forecast future performance. For example, Ruipérez-Valiente *et al.* [38] found that the best predictor was the grade achieved in the completed assignments when predicting dropout in a MOOC. In addition, Pigeau *et al.* [25] found that marks in quizzes where the best predictors when predicting success (i.e., who passes the course) in 12 MOOCs on programming languages, project management and startup creation.

Furthermore, the analysis of the HKUST MOOC corroborates that forum variables are not strong predictors. A possible reason can be because learners posted very few messages (42% of contributors in the HKUST MOOC only posted one message and 68% in the UC3M MOOC). Although results could differ in other contexts (such as in the article by Klüsener and Fortenbacher [35], where they obtained good accuracies when predicting success with and features such as the number of up-votes), this result suggests that forum variables may not be useful for predictions unless specific conditions are met. For example, it could be interesting to further analyze this issue in a MOOC where involvement in the forum was crucial and also analyze the effect of passive participation (e.g., users who read what others have done, but never post messages).

In addition, results showed that the predictive power in the first weeks was worse, probably because there were less data available and there were many learners who were exploring the course at initial stages but they were not committed to do the course. This lack of commitment was also observed with a decrease of engagement (i.e., activity) in videos and exercises over time, and it was also reported in other contributions (e.g., [24], [28]). Moreover, it was found that this decrease of engagement over weeks was found to be similar in both MOOCs (there was an exponential decrease in both MOOCs) despite the course duration was different. In contrast, a difference between both analyzed MOOCs was that the non-cumulative mode was not always the best in the HKUST MOOC, which suggests that non-cumulative mode could be limited if there are very few exercises in each module, as in this MOOC.

In terms of predictive power, results obtained were similar to those obtained in other contributions. For example, Pérez-Sanagustín *et al.* [40] obtained an RMSE of 0.18 when using all the interactions and RF to predict the overall grade, which is similar to the performance obtained here (best RMSE in the HKUST MOOC is 0.16 for FE). Moreover, Elbadrawy *et al.* [29] obtained RMSEs between 0.15 and 0.23 when predicting assignment grades in a MOOC using regression models (although no information is known about the kind of assignments and course methodology). This show that the values obtained here are similar to those obtained in other works. It may be possible to improve the models since our work does not focus on obtaining the best tuned model, as Ding *et al.* [28] did when incorporated more complex models using neural networks to predict grades of each chapter (their RMSEs improved from 0.25-0.40 to 0.12-0.15). However, the results presented here can be considered accurate enough so as to be used in real contexts and provide predictions that accurately indicate the range of grade (e.g., between 70-80%) the student is going to achieve.

In addition, a sensitivity analysis also showed that predictive models seemed to be robust. We measured how a percent increase of each variable (20%, as used by Hamby [53]) affected the prediction of grade of each assignment (using Model H, which contains all variables and RF, which is one of the most consistent algorithms in the analysis). Results showed that these variations in the input produced variations in the output smaller than 0.01 (and only 0.01 with *per_vtotal* in L1 and *avg_rep* in PS), which barely affect results. Moreover, the rate between the percentage of variation in output and input was low in most of the cases, which implies that models are barely affected by the uncertainty in the input.

The analysis also concluded that clickstream data (in this case including information about interactions with videos and access to the platform) could not considerably improve the predictive power when variables related to exercises were available (e.g., average grade or previous summative grades). However, clickstream data were useful to predict when information about exercises was missing (assuming that we exclude events about exercises from the clickstream category, although they are usually available in the clickstream and they could fall into this category). This suggests that researchers can develop accurate predictive models whenever variables related to exercises are available (as in the courseware data of edX) and in this case, other data limitations may be alleviated. Nevertheless, in the absence of variables related to exercises, it can be possible to develop acceptable models, as Alamri *et al.* [34] did, where not all weeks had quizzes to be used in the predictions, and they achieved accurate predictions of dropout based on variables related to activity (which are obtained here through clickstream).

With regard to the analysis of FE, results showed that the predictive power can vary depending on the type of questions, as also happens with the format of assignments (with concept-oriented assignments easier to predict than lab-oriented ones) and happened in a previous work [5]. In this case, MC are easier to predict than CD, which focus more on the assessment of skill. Furthermore, results showed that predictions of the FE and the FG were very related, and predictions of the FE could be used to predict FG without losing much predictive power. However, results showed that FG was easier to predict than the FE, which entails that the average performance in the long run may be easier to predict.

Finally, despite obtaining several conclusions about the factors affecting predictions, it is important to discuss about

the applicability and how they can make impact on learning. In this case, as the MOOC is run every year, it could be possible to use predictive models in consecutive editions to report instructors and students about their expected performance. However, as predictions can sometimes fail, they could be accompanied by the values of the most important indicators (as they are also obtained here) to give some explanations related to the predictions. In addition, for the pass/fail classification, a probability should be given instead of a binary result to be more precise. Finally, it would be important to involve instructors so they decide which interventions can be done and how to transmit prediction results to learners so they do not lose motivation, but they can reflect on their achievements and how to improve.

VI. CONCLUSION

This paper has analyzed several factors affecting predictions of grades in a MOOC. Results showed that the best variables were those related to exercises, and forum variables were not useful to predict. Clickstream data was found to be acceptable predictors when exercises variables were not available, but they did not enhance the predictive power when the latter were present. Predictive power was also better for concept-oriented assignments and best models usually contained only the last interactions. In addition, results showed that multiple-choice questions were easier to predict than coding questions, and the final exam grade was harder to predict than the final grade, based on different assignments during the course.

Despite these findings, there are some limitations that are worth mentioning. In this study, two MOOCs on the same topic but using different pedagogical approaches are compared. However, it would be relevant to analyze more courses to get further conclusions about the generalizability of the findings. In addition, there are some limitations related to the methodology. First, one limitation is about the way to filter students. Although this is a common limitation in MOOCs, where there are many learners who do not interact with the platform and need to be removed, there can be different criteria and they could affect the results. For example, instead of filtering out those who did not participate in the forum, we could have filtered those students who had not completed any assignment. Second, we gathered the variables at the end of this week as we were more focused on the factors that affect prediction than early prediction. For an analysis of early prediction, it would be better to stop the data collection some days before the deadline. Third, the analysis is based on a defined set of features, but other features could have been defined (e.g., navigation patterns, variables about self-regulation) and these features could also be relevant in the analysis. Fourth, the analysis is carried out with certain algorithms. More algorithms and more complex ones, such as neural networks, could be used to analyze whether they can improve the predictive power.

As future work, some other factors can be analyzed to get insight about whether or not they can have an influence on the predictive power. Some of these factors can be the use of new variables, the thematic area of the course, the evaluation system (e.g., analyzing differences between a mid-term and final exam grade), etc. In addition, it will be interesting to incorporate more courses with more different settings to have a better validation of the results. Moreover, it will be relevant to analyze how other prediction outcomes can be related, for example, drop out and final grade, etc. In addition, more research can be done on the type of exam questions, and their difficulty could be also analyzed to see if it also affects the predictive power. Finally, it would be important to develop systems that make use of these predictions in order to inform instructors and learners so that predictions can actually contribute to the improvement of success in MOOCs.

TABLE 6. Results of the predictive models in the HKUST MOOC using mae as metric.

Method		L1 W1	L2 W2	L3 W3	L4 W5	L5 W7	L6 W8	PS W9	PR W9	FE W_{10}
Mod. A	RG	.23	.23	.17	.19	.15	.17	.15	.14	.09
	RF	.20	.23	.16	.19	.13	.16	.15	.14	.09
	SVM	**.17**	.17	.12	.15	.11	.14	.13	.13	.08
	DT	.21	.24	.17	.19	.14	.16	.15	.14	.09
Mod. B	RG	.23	.22	.16	.17	.14	.15	.14	.14	.09
	RF	.20	.20	.14	.13	.11	.13	.13	.13	.08
	SVM	**.17**	**.16**	**.11**	**.10**	**.08**	**.10**	**.12**	**.12**	**.07**
	DT	.21	.21	.14	.13	.11	.13	.14	.14	.09
Mod. C	RG	.33	.41	.45	.41	.37	.29	.32	.25	.35
	RF	.33	.41	.45	.41	.37	.30	.31	.25	.34
	SVM	.22	.37	.38	.31	.26	.21	.21	.18	.26
	DT	.32	.41	.45	.42	.37	.29	.32	.25	.35
Mod. D	RG	.23	.23	.17	.19	.14	.17	.15	.14	.09
	RF	.20	.23	.17	.20	.13	.16	.15	.14	.09
	SVM	**.17**	.17	.13	.16	.11	.16	.13	.14	.08
	DT	.20	.24	.18	.20	.14	.16	.16	.15	.09
Mod. E	RG	.23	.22	.16	.17	.14	.16	.14	.14	.09
	RF	.20	.21	.14	.14	.11	.13	.13	.13	.08
	SVM	**.17**	.17	.12	.11	.09	.12	**.12**	.13	**.07**
	DT	.20	.20	.15	.14	.11	.13	.14	.14	.09
Mod. F	RG	.23	.26	.24	.23	.22	.24	.22	.19	.16
	RF	.20	.24	.19	.17	.13	.15	.16	.14	.09
	SVM	**.17**	.18	.16	.14	.13	.15	.15	.14	.10
	DT	.21	.25	.20	.18	.14	.15	.17	.17	.11
Mod. G	RG	.28	.30	.30	.24	.21	.20	.17	.16	.11
	RF	.27	.29	.28	.24	.20	.20	.16	.15	.11
	SVM	.20	.22	.24	.20	.17	.18	.15	.16	.09
	DT	.28	.31	.29	.25	.21	.21	.18	.17	.13
Mod. H	RG	.23	.22	.16	.17	.14	.15	.14	.14	.09
	RF	.20	.21	.14	.14	.11	.13	.14	.13	.08
	SVM	.18	.17	.12	.12	.10	.13	.13	.13	.08
	DT	.20	.22	.14	.15	.11	.13	.15	.15	.09

Note: Best MAE values per column are in bold

APPENDIX

This appendix aims to report the results in other metrics. These other metrics may not be the recommended ones for this context [49], although they are commonly used in the literature, so this appendix allows researchers comparing their results with those presented in this paper with different metrics. Particularly, prediction of grades is also reported using MAE and prediction of pass/fail are also presented using F1 score, which is a measure which takes into account both recall and precision.

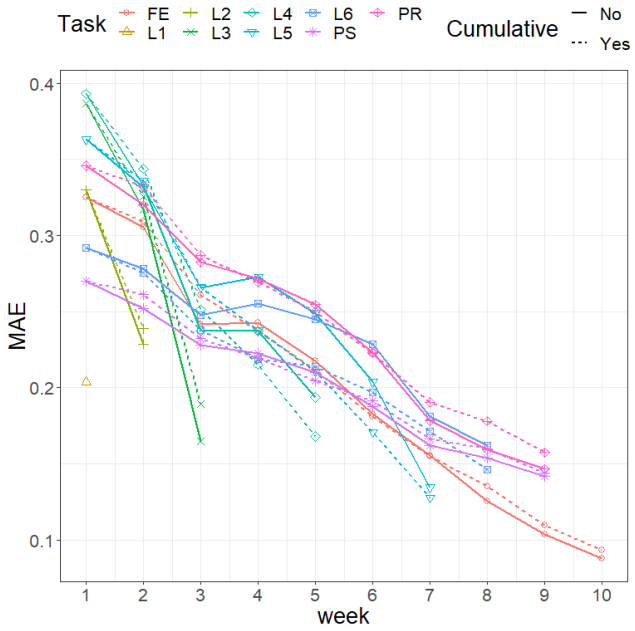

FIGURE 9. Prediction of grades using data from previous topics in the HKUST MOOC.

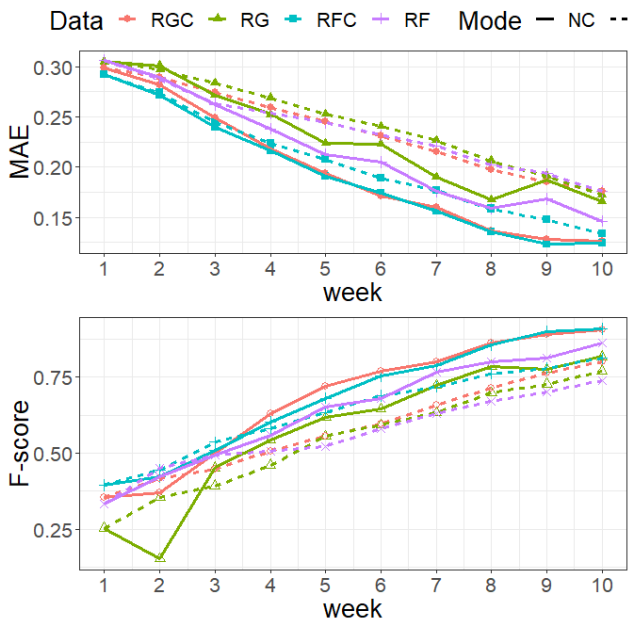

FIGURE 10. Prediction of the FG with RG and RF with (RGC and RFC) and without using clickstream data (RG and RF) in cumulative (C) and non-cumulative (NC) mode.

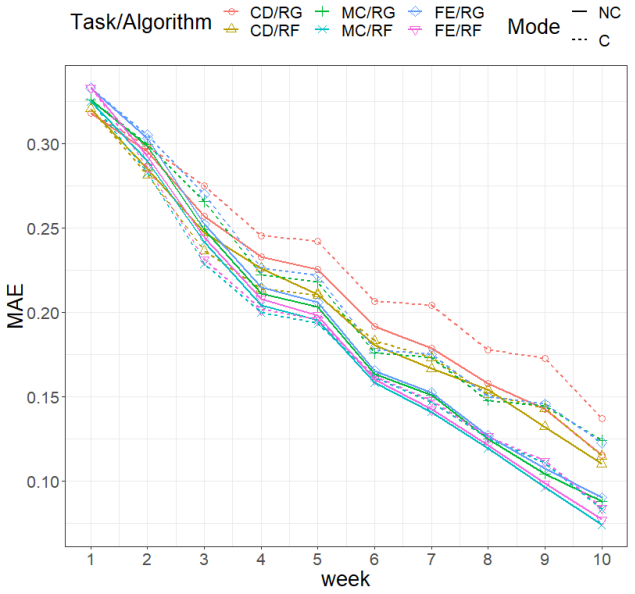

FIGURE 11. Evolution of the predictive power when predicting MC, CD, and FE itself, using cumulative (C) and non-cumulative (NC) mode.

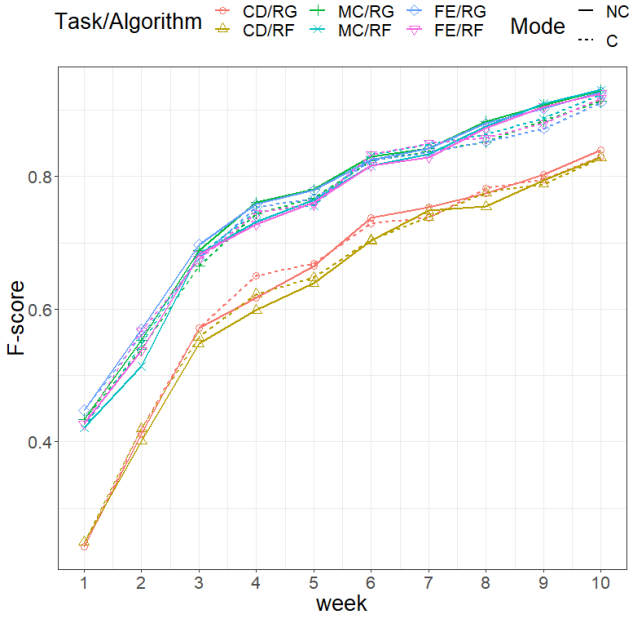

FIGURE 12. Evolution of the predictive power when predicting MC, CD, and FE itself, using cumulative (C) and non-cumulative (NC) mode.

Table 6 shows that although MAE is used the conclusions are very similar to those obtained with RMSE. It can be shown that previous grades enhance the predictive power (as Model B is the best model), forum variables are not useful to predict, non-cumulative mode is usually better than cumulative mode (although not always, as happened when analyzing with RMSE). In addition, clickstream variables are found not to enhance predictive models and they are not enough unless combined with other variables (e.g., exercises). However, the main difference is that SVM is the best algorithm in terms of MAE. The evolution of the predictive power shows a similar trend using MAE, as shown in Fig. 9.

With regard to the prediction of FG, results (Fig. 10) also show that clickstream can be useful to predict when variables related to exercises are not available, and both MAE and F-score improves when clickstream variables are added. These conclusions were the same as those presented in RQ2.

As for the predictive power of MC and CD, Fig. 11 and 12 present the evolution of the predictive power using MAE

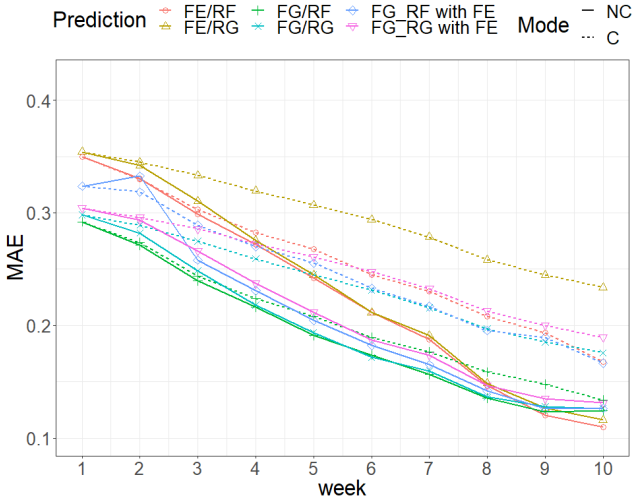

FIGURE 13. Comparisons of predictive models to forecast FE exam and FG.

and F-score, respectively. The figure shows that MC is easier to predict and the predictive power achieves are very good (MAE is below 0.1, which means that the errors are in average of only some tenths).

Finally, Fig. 13 presents the comparison between the predictive power of the FG and the FE. In that figure, it can also be observed that FE is generally harder to predict, although in this case, the predictive power is very similar (even better for the FE) when all (or almost all) interactions are available (weeks 9 and 10). This final improvement was also present in Fig. 8 although RMSE was better for FG prediction. This result serves to make note that while FE is harder to predict, it can be as easy as the FG when the exam is near (weeks 9 and 10).

ACKNOWLEDGMENT

This publication reflects the views only of the authors and funders cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] S. Joksimović, O. Poquet, V. Kovanović, N. Dowell, C. Mills, D. Gašević, S. Dawson, A. C. Graesser, and C. Brooks, "How do we model learning at scale? A systematic review of research on MOOCs," *Rev. Educ. Res.*, vol. 88, no. 1, pp. 43–86, Feb. 2018.

[2] R. F. Kizilcec and G. L. Cohen, "Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 17, pp. 4348–4353, Apr. 2017.

[3] W. Xing and D. Du, "Dropout prediction in MOOCs: Using deep learning for personalized intervention," *J. Educ. Comput. Res.*, vol. 57, no. 3, pp. 547–570, Jun. 2019.

[4] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *Internet Higher Edu.*, vol. 28, pp. 68–84, Jan. 2016.

[5] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, and C. Delgado Kloos, "Analysing the predictive power for anticipating assignment grades in a Massive Open Online Course," *Behav. Inf. Technol.*, vol. 37, nos. 10–11, pp. 1021–1036, Apr. 2018.

[6] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, I. Estevez-Ayres, and C. D. Kloos, "A learning analytics methodology for understanding social interactions in MOOCs," *IEEE Trans. Learn. Technol.*, vol. 12, no. 4, pp. 442–455, Oct. 2019.

[7] Y. S. Tsai and D. Gasevic, "Learning analytics in higher education—Challenges and policies: A review of eight learning analytics policies," in *Proc. LAK*, Vancouver, BC, Canada, 2017, pp. 233–242.

[8] A. Cano and J. D. Leonard, "Interpretable multiview early warning system adapted to underrepresented student populations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 198–211, Apr. 2019.

[9] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, and C. D. Kloos, "Analyzing students' persistence using an event-based model," in *Proc. LASI*, Vigo, Spain, 2019, pp. 56–70.

[10] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions," *IEEE Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, Jul. 2019.

[11] A. F. Cabrera, A. Nora, and M. B. Castaneda, "College persistence: Structural equations modeling test of an integrated model of student retention," *J. Higher Edu.*, vol. 64, no. 2, p. 123, Mar. 1993.

[12] K. Casey and P. Gibson, "Mining moodle to understand student behavior," in *Proc. ICEP*, Maynooth, Ireland, 2010.

[13] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proc. EDM*, Córdoba, Spain, 2009, pp. 41–50.

[14] J. G. Glynn, P. L. Sauer, and T. E. Miller, "A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen," *Int. Bus. Econ. Res. J.*, vol. 1, no. 8, pp. 79–86, 2000.

[15] S. Herzog, "Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen," *Res. High Educ.*, vol. 46, no. 8, pp. 883–928, Dec. 2005.

[16] B. C. Christensen, B. Bemman, H. Knoche, and R. Gade, "Identifying students struggling in courses by analyzing exam grades, self-reported measures and study activities," in *Proc. SLERD*, Aalborg, Denmark, 2018, pp. 167–176.

[17] B. C. Christensen, B. Bemman, H. Knoche, and R. Gade, "Pass or fail? Prediction of students' exam outcomes from self-reported measures and study activities," *Interact. Des. Archit.*, vol. 39, no. 17, pp. 44–60, 2019.

[18] A. M. Shahiri and W. Husain, "A review on predicting students' performance using data mining techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, Dec. 2015.

[19] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," *IEEE Trans. Learn. Technol.*, vol. 7, no. 4, pp. 346–359, Oct. 2014.

[20] M. L. Bote-Lorenzo and E. Gómez-Sánchez, "An approach to build *in situ* models for the predictionof the decrease of academic engagement indicators in massive open online courses," *J. Universal Comput. Sci.*, vol. 24, no. 8, pp. 1052–1071, Aug. 2018.

[21] G. Chen, D. Davis, C. Hauff, and G.-J. Houben, "On the impact of personality in massive open online learning," in *Proc. Conf. User Model. Adaptation Personalization (UMAP)*, Halifax, NS, Canada, 2016, pp. 121–130.

[22] S. Kolowich. (2013). Coursera Takes a Nunaced View of MOOC Dropout Rates. The Chronicle of Higher Education. [Online]. Available: https://www.chronicle.com/blogs/wiredcampus/coursera-takes-a-nuanced-view-of-mooc-dropout-rates/43341

[23] N. Wu, L. Zhang, Y. Gao, M. Zhang, X. Sun, and J. Feng, "CLMS–Net: Dropout prediction in MOOCs with deep learning," in *Proc. ACM Turing Celebration Conf. ACM TURC*, 2019, Art. no. 75.

[24] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods," in *Proc. EMNLP Workshop Anal. Large Scale Social Interact. MOOCs*, Doha, Qatar, 2014, pp. 60–65.

[25] A. Pigeau, O. Aubert, and Y. Prié, "Success prediction in MOOCs a case study," in *Proc. EDM*, Montreal, QC, Canada, 2019, pp. 390–395.

[26] B. Xu and D. Yang, "Motivation classification and grade prediction for MOOCs learners," *Comput. Intell. Neurosci.*, vol. 2016, no. 4, pp. 1–7, Jan. 2016.

[27] X. Li, T. Wang, and H. Wang, "Exploring N-gram features in clickstream data for MOOC learning achievement prediction," in *Proc. DASFAA*, Suzhou, China, 2017, pp. 328–339.

[28] M. Ding, K. Yang, D.-Y. Yeung, and T.-C. Pong, "Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses," in *Proc. 9th Int. Conf. Learn. Anal. Knowl. (LAK)*, Tempe, Arizona, 2019, pp. 135–144.

[29] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Computer*, vol. 49, no. 4, pp. 61–69, Apr. 2016.
[30] Á. Pérez-Lemonche, G. Martínez-Muñoz, and E. Pulido-Cañabate, "Analysing event transitions to discover student roles and predict grades in MOOCs," in *Proc. ICANN*, 2017, pp. 224–232.
[31] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models," in *Proc. EDM*, Raleigh, NC, USA, 2016, pp. 484–489.
[32] Z. A. Pardos, Y. Bergner, D. T. Seaton, and D. E. Pritchard, "Adapting Bayesian knowledge tracing to a massive open online course in edX," in *Proc. EDM*, Memphis, TN, USA, 2013, pp. 137–144.
[33] J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. Delgado-Kloos, "Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning," in *Proc. EC-TEL*, Leeds, U.K., 2018, pp. 355–369.
[34] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, and C. Stewart, "Predicting MOOCs dropout using only two easily obtainable features from the first week's activities," in *Proc. ITS*, Kingston, Jamaica, 2019, pp. 163–173.
[35] M. Klusener and A. Fortenbacher, "Predicting students' success based on forum activities in MOOCs," in *Proc. IEEE 8th Int. Conf. Intell. Data Acquisition Adv. Comput. Syst., Technol. Appl. (IDAACS)*, Warsaw, Poland, Sep. 2015, pp. 925–928.
[36] C. Taylor, K. Veeramachaneni, and U.-M. O'Reilly, "Likely to stop? Predicting stopout in massive open online courses," Aug. 2014, *arXiv:1408.3382*. [Online]. Available: https://arxiv.org/abs/1408.3382
[37] C. Brooks, C. Thompson, and S. Teasley, "Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs)," in *Proc. 2nd ACM Conf. Learn. Scale*, Vancouver, BC, Canada, 2015, pp. 245–248.
[38] J. A. Ruipérez-Valiente, R. Cobos, P. J. Muñoz-Merino, Á. Andujar, and C. Delgado Kloos, "Early prediction and variable importance of certificate accomplishment in a MOOC," in *Proc. EMOOCs*, 2017, pp. 263–272.
[39] T.-Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior–based grade prediction for MOOCs via time series neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 5, pp. 716–728, May 2017.
[40] M. Pérez-Sanagustín, K. Sharma, R. Pérez-Álvarez, J. Maldonado-Mahauad, and J. Broisin, "Analyzing learners' behavior beyond the MOOC: An exploratory study," in *Proc. EC-TEL*, Delft, The Netherlands, 2019, pp. 40–54.
[41] C. G. Brinton and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 2299–2307.
[42] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue, "Modeling and predicting learning behavior in MOOCs," in *Proc. 9th ACM Int. Conf. Web Search Data Mining (WSDM)*, San Francisco, CA, USA, 2016, pp. 93–102.
[43] P. M. Moreno-Marcos, T. De Laet, P. J. Muñoz-Merino, C. Van Soom, T. Broos, K. Verbert, and C. Delgado Kloos, "Generalizing predictive models of admission test success based on online interactions," *Sustainability*, vol. 11, no. 18, p. 4940, Sep. 2019.
[44] C. Romero and S. Ventura, "Guest editorial: Special issue on early prediction and supporting of learning performance," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 145–147, Apr. 2019.
[45] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in *Proc. AIED*, Madrid, Spain, 2015, pp. 54–63.
[46] R. F. Kizilcec and S. Halawa, "Attrition and achievement gaps in online learning," in *Proc. L@S*, Vancouver, BC, Canada, 2015, pp. 57–66.
[47] J. Gardner, C. Brooks, J. M. Andres, and R. Baker, "Replicating MOOC predictive models at scale," in *Proc. 5th Annu. ACM Conf. Learn. Scale (L@S)*, London, U.K., 2018, Art. no. 1.
[48] edX. (2018). *EdX Research Guide Release*. [Online]. Available: https://buildmedia.readthedocs.org/media/pdf/devdata/latest/devdata.pdf
[49] R. Pelánek, "Metrics for evaluation of student models," *J. Educ. Data Mining*, vol. 7, no. 2, pp. 1–19, Jun. 2015.
[50] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interact. (ACII)*, Geneva, Switzerland, Sep. 2013, pp. 245–251.
[51] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales, and J. Munoz-Gama, "Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in massive open online courses," *Comput. Human Behav.*, vol. 80, pp. 179–196, Mar. 2018.
[52] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Proc. NIPS*, Lake Tahoe, NV, USA, 2013, pp. 431–439.
[53] D. M. Hamby, "A comparison of sensitivity analysis techniques," *Health Phys.*, vol. 68, no. 2, pp. 195–204, Feb. 1995.

PEDRO MANUEL MORENO-MARCOS received the B.S. degree in telecommunications technologies engineering and the M.S. degree in telecommunication engineering and telematics engineering from the Universidad Carlos III de Madrid, Leganés, Spain, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree in telematics engineering.

His thesis focuses on the prediction of learners' behaviors and indicators in online learning platforms. From 2016 to 2017, he was with the Department of Telematics, Universidad Carlos III de Madrid, with a collaboration internship, where he became a Training Teacher with a fellowship from the Ministerio de Ciencia, Innovación y Universidades, in 2017. He currently holds this position. His research interests include learning analytics, educational data mining, and massive open online courses (MOOCs).

Mr. Moreno-Marcos awards and honors include the Ubica Award for the Best Academic Achievement, the Best Master's Thesis Award by the Spanish Chapter of the IEEE Education Society, the Fundación Telefónica Award for the Best Master's Thesis in Technology and Education, and extraordinary awards as the best student in the bachelor's and master's degree, among others.

TING-CHUEN PONG received the M.S. and Ph.D. degrees in computer science from Virginia Polytechnic Institute and State University, in 1984.

He is a founding Faculty Member of HKUST, where he had served as an Acting Provost, the Associate Vice-President of academic affairs, the Director of the Center for Engineering Education Innovation, and the Associate Dean of engineering. He was an Academic Research Advisor of the Hong Kong University Grants Committee (UGC), from 2010 to 2012. He is currently a member of the Quality Assurance Council of the UGC and the Chairman of the Qualifications and Accreditation Committee of the Hong Kong Council for Accreditation of Academic and Vocational Qualifications. Before joining HKUST, he was an Associate Professor of computer science with the University of Minnesota. He is currently a Senior Advisor to the Provost, the Acting Director of the Center for Education Innovation, and a Professor of computer science and engineering with Hong Kong University of Science and Technology (HKUST). His research interests include computer vision, multimedia computing, and IT in education.

Dr. Pong was a recipient of the HKUST Excellence in Teaching Innovation Award, in 2001. In 2014, he led the HKUST Team of the Wharton-QS Stars Awards Competition. He was selected as the Winner of the Natural Sciences Award and the Runner-Up of the Hybrid Learning Award.

PEDRO J. MUÑOZ-MERINO (Senior Member, IEEE) was born in Cuenca, Spain, in 1979. He received the Telecommunications Engineering degree from the Universidad Politécnica de Valencia, Spain, in 2003, and the Ph.D. degree in telematics engineering from the Universidad Carlos III de Madrid, Spain, in 2009.

His skills and experience include research and development in learning analytics, educational data mining, the evaluation of learning experiences, user studies, gamification, or intelligent tutoring systems. He is an Associate Professor with the Universidad Carlos III de Madrid. He is the Coordinator of the LALA Project supported by the European Union for the adoption of learning analytics in Latin America. He has participated in more than 40 research projects at international and national levels, including also several contracts with companies, being the Principal Investigator of several of them related with learning analytics, educational data mining, and adaptive systems. He is author of more than 120 scientific publications including more than 35 in journals indexed in the JCR. He has also coordinated the development and deployment of different learning analytics tools.

Dr. Muñoz-Merino has been a recipient of several awards such as the best master thesis in telematics engineering, a special award for the Ph.D. degree, and several nominations for the best paper award in conferences or several recognitions as a supervisor of the best master thesis or Ph.D. degree awards. He is currently the coordinator of the work group about learning analytics and adaptive systems of the IEEE Spain chapter about education.

CARLOS DELGADO KLOOS (Senior Member, IEEE) received the Ph.D. degree in computer science from the Technical University of Munich and the Ph.D. degree in telecommunications engineering from the Technical University of Madrid, in 1986.

His main research focus is in educational technologies including MOOCs, learning analytics, augmented reality, gamification, and mobile learning. Since 1996, he has been a Full Professor of telematics engineering with the Universidad Carlos III de Madrid, where he is the Director of the UNESCO Chair on Scalable Digital Education for All and the GAST Research Group. He is also Vice-President of strategy and digital education. He coordinates and has coordinated many projects related to educational technologies at international, national, and regional levels, and contracts with companies. He is the author of more than 400 scientific articles, including more than 80 in JCR-indexed journals. He has been a recipient of several awards related to educational technologies. He is often invited as a keynote speaker at different conferences and events.