Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Delgado Kloos, C., Prediction in MOOCs: A review and future research directions, ***IEEE Transactions on Learning Technologies***, July 2018

DOI: 10.1109/TLT.2018.2856808

Prediction in MOOCs: A review and future research directions

Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino, *Senior Member, IEEE* and Carlos Delgado Kloos, *Senior Member, IEEE*

Abstract—. This paper surveys the state of the art on prediction in MOOCs through a Systematic Literature Review (SLR). The main objectives are: (1) to identify the characteristics of the MOOCs used for prediction, (2) to describe the prediction outcomes, (3) to classify the prediction features, (4) to determine the techniques used to predict the variables, and (5) to identify the metrics used to evaluate the predictive models. Results show there is strong interest in predicting drop-outs in MOOCs. A variety of predictive models are used, though regression and Support Vector Machines stand out. There is also wide variety in the choice of prediction features, but clickstream data about platform use stands out. Future research should focus on developing and applying predictive models that can be used in more heterogeneous contexts (in terms of platforms, thematic areas, and course durations), on predicting new outcomes and making connections among them (e.g., predicting learners' expectancies), on enhancing the predictive power of current models by improving algorithms or adding novel higher-order features (e.g., efficiency, constancy, etc.).

Index Terms—Discussion forums, Distance Learning, Learning Environments, Machine Learning

1 Introduction

Massive Open Online Courses (commonly known by the acronym MOOCs) are open courses designed to provide educational content to a large number of participants through an online platform, and with free access [1]. The term was coined in 2008, and has become more and more popular since 2012, creating a new paradigm in education. MOOCs originally appeared to enable learners all over the world to gain access to introductory courses from universities. However, currently there are MOOCs about many different thematic areas and levels, and as Christensen et al. [2] have shown, the target learners are not only undergraduate students anymore: Anyone with an interest can take one of these courses.

One of the main characteristics of MOOCs is the large number of enrollees, due to the nature of these courses (there can be typically thousands of enrollees and the fact that these courses are open contributes to getting these numbers). This allows a vast amount of information to be collected about what is happening in the course for further analysis. Interestingly, most platforms store large amounts of data from all the interactions learners have with course contents. These interactions include, among others, course navigation events, video events (i.e., when a user plays a video, changes the speed, and so on) and logs related to the exercises (i.e., the number of attempts and scores for each exercise, number of hints used, and so on).

- P.M. Moreno-Marcos is with the *Universidad Carlos III de Madrid, Leganés, ES 28911*. E-mail: *pemoreno@it.uc3m.es*.
- C. Alario-Hoyos is with the *Universidad Carlos III de Madrid, Leganés, ES 28911*. E-mail: *calario@it.uc3m.es*.
- P.J. Muñoz-Merino is with the *Universidad Carlos III de Madrid, Leganés, ES 28911*. E-mail: *pedmume@it.uc3m.es*.
- C. Delgado Kloos is with the *Universidad Carlos III de Madrid, Leganés*, ES 28911. E-mail: *cdk@it.uc3m.es*.

Apart from that, forum posts in MOOCs can also be collected and analyzed, as they provide relevant information not only about users' attitudes or sentiments (i.e., a user can show what his performance or engagement is like by the comments he posts), but also about social interactions. In this area, it is possible to analyze some metrics related to the network formed from the interactions of learners in the course (e.g., degree, centrality, etc.). Other interesting related aspects that can be considered are learners' reputation and their sense of community [3].

All these data can be used not only to detect problems from the observed data, but also to predict learners' behaviors and learning outcomes. These predictions can be very helpful for the different stakeholders for several reasons. Teachers can anticipate possible problems with learners and adapt the course or methodology to enhance the learning experience. In addition, instructors or the institution itself can use the predictions to make decisions about the curricular design and to carry out interventions. Furthermore, learners can receive information about their learning process that will enable them to reflect on how they are doing and improve their performance.

The first question when predicting based on events related to users in a MOOC is what outcome is going to be predicted. One traditional example is the dropout rate. As MOOCs are open, many people enroll in these courses without really thinking about the work involved, and, eventually, they end up withdrawing. Gütl et al. [4] compared dropout rates between face-to-face courses and online courses and stated that they can be 10 % or 20 % higher when courses are online. For this reason, there have been several contributions trying to predict dropout rates in MOOCs [5], [6]. Apart from dropouts, grade prediction, engagement and sentiments are other features

Published by the IEEE Computer Society

that have been predicted in the literature about MOOCs [7], [8].

In the literature, there are many articles related to prediction in MOOCs, but as there can be a large list of variables to forecast with different indicators and techniques, there is not a clear vision about what has been researched in this area to be able to innovate with new contributions. Although each article evaluates prediction techniques to use the most appropriate ones for its case and uses available data from one or several MOOCs, not all types of variables are used in all contexts. Therefore, it is interesting to identify which indicators are more frequent in the literature (because of their effectiveness), what problems have been addressed and which ones offer new possibilities for research, and what predictive models achieve accurate results for each context.

This work conducts a Systematic Literature Review (SLR) of prediction in MOOCs. The contributions in the state of the art are structured according to five main research questions, which cover the different characteristics mentioned previously. The ultimate goal is to be able to detect future challenges and possible future research directions in the area of prediction in MOOCs.

RQ1: What are the most common characteristics of the MOOCs that have been used for prediction?

RQ2: What outcomes have been predicted in contributions about MOOCs?

RQ3: What are the prediction features that are used to build prediction models in MOOCs?

RQ4: What are the techniques/models used for prediction in MOOCs?

RQ5: What metrics have been used to evaluate prediction results in MOOCs?

The structure of the paper is as follows. Section 2 provides a background on what has been researched about prediction in education in general and then discusses the particularities for the specific case of MOOCs. Section 3 describes the systematic methodology that has been used to collect and select the articles, which provide the necessary information to answer the research questions. Results are discussed in Section 4, while the future research directions are provided and justified in Section 5. Finally, the main conclusions are described in Section 6.

2 Related Work

2.1 Prediction in education

Prediction is a wide area of study and there have been contributions in many different fields, including education. Currently, most of the work in the literature on prediction in education refers to "closed" courses in schools, high schools or universities. In general, predictive models are being used for both anticipating future events (i.e., forecasting students' performance at the end of the course) and for detecting patterns in learners' behaviors (i.e., engagement, participation in forums, behavior changes, and so on). This section will outline relevant outcomes obtained in this field from the literature, taking into account the two possible approaches.

First of all, one of the main interests has been in predicting whether or not a student will finish and pass the course. In order to do that, most researchers have used the cumulative grade point average (CGPA) [9], [10]. Demographic data (age, gender, mother tongue, origin, etc.) are also commonly used, as well as other information related to platform logs; for example, Gašević et al. [11] used Moodle logs and submission results for predicting learners' performance. Combining data from different sources can also contribute to improving prediction outcomes; for example, Aguiar et al. [12] found that on many occasions, student performance features (like the CGPA) were not enough to predict dropouts, and engagement variables needed to be added (e.g., number of accesses to the platform) to achieve good accuracy results. Timing is also a key factor that can affect the results of the experiments. Lykourentzou et al. [13] performed a study to predict dropouts in two introductory courses from the University of Athens. They considered only demographic variables at the beginning and they added data from the course as the course evolved. They found that while the overall classification rate was between 75 and 85 % from the first section of the course, they could reach a 97–100 % rate in the final phases.

Instead of predicting whether a learner will pass a single course or not, one could try to predict whether a student will complete a whole degree. This was the case with Daud et al. [14], who took data from students from several Pakistani universities for predicting this, using variables related to family expenditure, family income, personal information and students' family assets. Nevertheless, prediction models can be extended with data about students' performance during the degree, or at specific moments. For example, Neumann, Neumann, and Lewis [15] predicted completion in a master's degree using data from the first (and crucial) course of the program.

When using prediction techniques in face-to-face courses, students' grades are an important outcome. Several contributions analyzed different items that could correlate with grades. For example, Meier et al. [16] showed that in-class exam results are better predictors of performance than assignments, but, in contrast, Huang and Fang [17] researched the effect of prerequisite grades and midterm exams, and found that prerequisite grades are not better indicators than the average grade of the whole degree (CGPA) and that the first midterm exam is not reliable at all when making predictions. As face-to-face courses enable more interactions with classmates and teamwork, peer grading has also been considered for forecasting grades. However, Sajjadi, Alamgir, and von Luxburg [18] showed that the high variance among grades and the wide-ranging sources of grading errors suggest that peer grades are not reliable. In this case, timing considerations are also relevant since prediction purposes can vary over a number of weeks. For example, at the beginning of the course, one interest may be forecasting the midterm results, whereas in the last week the aim of prediction is to know the final exam result (or the final grade). In this case, as the course evolves and more

information becomes available, results can improve in the final weeks [19]. This was also corroborated by Okubo et al. [20], who predicted grades using neural networks, and improved the accuracy from 50 % in the first week to 100 % in the tenth.

Another common goal in works on prediction is identifying students at risk of failing a course with a view to performing interventions with the aim of encouraging students to work harder, and have a greater chance of passing the course. In this case, Marbouti, Diefes-Dux, and Madhavan [21] highlighted the importance of minimizing false negatives to ensure that students at risk are warned about their situation. Generally, students may or may not be at risk, but other approaches have considered more states with a view to taking corrective measures, such as performance at or above (or below) the course mean, or trending towards underperformance [22].

Student behavior is another major area of study in prediction, and forum posts are an important source of information to be considered (if available, since traditional face-to-face courses do not normally collect social interactions). Romero et al. [23] took forum messages and classified posts according to their relationship with the subject and the knowledge students demonstrate with their posts. Additionally, these authors considered social measures, such as the degree of centrality or prestige of the student. In the social area, Chen, Vorvoreanu, and Madhavan [24] gathered tweets using geolocation to retrieve only information within a radius of 1.3 miles from the campus, and were able to identify students' problems, such as a heavy study load, a lack of social engagement, negative emotions, or sleep problems.

Apart from the forum contributions, Xenos [25] developed a Bayesian Network (BN), which included information regarding different areas, including abilities, motivations, and tutor efforts, in order to model students' behaviors. This author concluded that the BN design is crucial for this type of analysis. However, in contrast to the general purpose, the aim was not predicting students' performance beforehand. Instead, he proposed a model to analyze data after the course, which can be another purpose of predictive models.

2.2 Why is it important to predict in MOOCs?

MOOCs have particularities, and a special focus will be placed on them when talking about prediction. At first sight, MOOCs are different from other courses where there are research works about prediction due to the vast amount of data provided by the huge number of learners. Apart from the volume, variety is another characteristic since users can be very different (in terms of culture, education, personality, and so on) and there can be many different behaviors. Additionally, there is an intensive use of videos and social interactions in the forum compared to other courses. This section will emphasize why MOOCs are different and will provide a background on the relevant work in the literature. The structure will be similar to that in the previous section; common prediction purposes, such as dropouts, scores prediction and forum variables, will be considered. It is worth noting that courses restricted to students on campus are excluded from the definition of MOOCs and, therefore, such "closed" courses will be beyond the scope of this study.

First of all, attrition is considered one of the main research issues in prediction in MOOCs. As courses can be accessed freely, it is very common for people to enroll just to browse through the content without any intention of finishing the course [26] (MOOCs may be considered as something optional for many learners). This entails a massive number of enrollees in MOOCs, but with a very low completion rate [27]. Thus, there is a need to optimize predictive models, not only for the case of learning outcomes (as in most formal education settings) but also for learner dropout in the course. Moreover, it is important to balance dropout outcomes (i.e., whether a learner is going to withdraw from the course or not) and learning outcomes. For example, Kizilcec, Bailenson, and Gomez [28] showed that the physical appearance of the instructor in MOOC videos (e.g., when the video shows the instructor's face) has little effect on learning outcomes but increases dropout for some learners.

In this field, Halawa, Greene, and Mitchell [29] analyzed learner activity features to forecast dropouts and found that absence times over three weeks have a clear relation with dropsouts; risk signals can be effective after only two inactive weeks and could be used for interventions. Among other prediction features, Taylor, Veeramachaneni, and O'Reilly [30] found a strong correlation between dropouts and problem submissions or features that involved inter-student collaboration (forum and wikis). They also showed the effectiveness of sophisticated features, such as the percentile of a student when compared to other students, or the lab grade over time. Moreover, Kizilcec and Halawa [31] identified learners at risk and invited them to complete a survey. Results showed that lack of time, course format, and difficulty were the three main reasons for dropping out of the course.

For the dropout problem, timing considerations are also significant since some analyses have shown that 75 % of dropouts occur in the first weeks of a MOOC [32] and, in general, people who start early are less likely to drop out [33]. For this reason, there have been several attempts to predict the point in a course at which a student will leave. Xing et al. [34] implemented a General Bayesian Network (GBN) and a decision tree with some predictors, such as the number of discussion posts, number of forum views, number of quiz views, number of module views, number of active days, and social network degree, with the aim of forecasting the dropout week. Kloft et al. [35] also used Support Vector Machines (SVMs) to predict dropouts over weeks and found that models outperformed baselines in advanced phases as there is not enough information at the beginning for prediction. Similarly, Cobos and Macías Palla [36] developed an interactive tool that included visualizations about the performance of different predictive models to predict certificate earners over time.

The last example differed from the previous ones because completing the course is not the same as passing it.

Because of that, as well as dropouts, the final score learners will get in the course is also worth predicting. In this case, MOOCs are also different as the assessment system may differ from that of face-to-face courses. In order to retrieve information for prediction, researchers may take indicators that are not available in other types of courses (e.g., information about video watching is not usually available in face-to-face courses). In MOOCs, prediction features are usually related to the platform use, forum activity, videos watched (number of videos watched/downloaded, events of play/pause) [37], and the results of the assignments. For example, Ren, Rangwala, and Johri [38] used these kinds of indicators to predict the score the learner was going to achieve in the following graded homework activity. Results showed that the best results were obtained in the middle of the course and the number of previous quizzes attempted before the graded one was the variable with the highest correlation. Moreover, Yang et al. [39] used a time series neural network to predict scores based on only previous assessment grades or their combination with clickstream data. They showed how video-watching clickstream could improve the predictive power. Sinha and Cassell [40] also created indicators about *burstiness* (they measured the number of video plays, chapters with interaction, or forum posts divided by the number of different days the learner interacted with the course) to separate learners' achievement into four categories (low, medium, high, and very high).

However, there are many other kinds of features (e.g., demographics) that have also been considered in the literature. For example, Kennedy et al. [41] considered the common types of variables, such as assignment submissions, assignment switches (i.e., the number of times a learner goes from one assignment to another), active days and total points, but also added special activities, such as graph coloring activities and those related to the knapsack problem, to predict the final score.

Forum posts are also an important focus in research and they can be a relevant source of prediction features (e.g., the number of posts, number of replies, number of votes up, and so on) that are usually not available in face-to-face courses. The information these posts provide could be processed through Natural Language Processing (NLP) techniques to build the predictive model, as Robinson et al. [42] did to forecast certificate earners in a six-week MOOC from edX. Furthermore, Chen et al. [43] developed a visualization tool with a predictive model to forecast dropout. They included the number of posts as a feature and also provided visualizations regarding the forum activity together with the dropout prediction results.

In addition, there have been several contributions that tried to classify messages posted in MOOC forums. Brinton et al. [44] analyzed 73 MOOCs from Coursera to classify posts according to their relevance using algorithms based on HITS (Hyperlink-Induced Topic Search) and TF-IDF (Term Frequency – Inverse Document Frequency). They also analyzed the decline of forum participation over time and found that while participation by the teaching staff increased the overall forum activity, in the long run it did not slow down the decline rate of messages over time. Ramesh et al. [45] also suggested a classification of posts according to different aspects (coarse aspects, fine aspects, sentiments, and sentiment toward online courses). In each category, several subtypes were defined to be more specific (e.g., fine aspects could be divided into lecture-video, lecture-audio, lecture-content, quiz-submission, quiz-grading, and so on). Moreover, Bakharia [46] worked on classifying forum posts according to three aspects (confusion, urgency, and sentiment) over three data sets.

Other approaches in the social area have tried to identify the people who contribute the most in the forums and their possible relationship with course completion [47], the participation of learners in peer reviews [48], and the personality of students, which was classified according to the Big Five personality dimensions [49] (openness, extraversion, conscientiousness, agreeableness, and neuroticism) by Chen et al. [50].

It is important to note that although there can be some common areas of interest in prediction, there were articles that took different approaches. For example, as in a MOOC there can be thousands of enrolled users and it would be very time-consuming for an instructor to reply to all posts in the forum, Chaturvedi et al. [7] applied logistic regression and decision trees to identify whether or not there would be intervention from an instructor in a message. Also, with forum posts, but with click patterns, Yang, Kraut, and Rose [51] developed a model to identify confusion among students (e.g., when a student says, "I'm stuck"). This can be very useful as in a MOOC it is more difficult to identify learners' difficulties than in a face-to-face course where the instructor sees the students regularly. Although teamwork between unknown people is not very common in MOOCs, Yang, Wen, and Rosé [52] took data from a NovoEd MOOC about *Constructive Classroom Conversations* where students needed to initiate or join a group at the beginning. These data were used to predict teamwork quality and to identify leaders in the group and the worst-performing learners, who are often a bottleneck for the team. Other examples of prediction outcomes include course satisfaction [53], student navigation in the course [54], and prerequisites between lectures [55], among others.

In this section, a general overview of prediction in education and particularly in MOOCs was presented. This paper goes beyond and provides an analysis of what has been researched, separating the characteristics of the studies: prediction outcomes, prediction features, techniques, course contexts, etc. This analysis is a novel contribution aimed at structuring the contributions of the current state of the art in prediction in MOOCs. The benefits of this work include the identification of what has been done in the past with a view to taking ideas and innovating. Furthermore, this contribution focuses on providing future research direction to help researchers choose the direction to take in their studies to advance in the state of the art.

3 METHODOLOGY

3.1 Description of the methodology

The methodology of this review follows the guidelines for a Systematic Literature Review (SLR) [56], [57]. This approach allows a wide coverage of results to be obtained that can be used for the analysis and discussion of the different articles that have been published in the literature. Other studies in the area of science education have used this methodology to elaborate their reviews (e.g. [58], [59], [60]).

The literature search was performed through two of the most common databases for retrieving scientific works: Scopus and ISI Web of Knowledge [61]. These databases include most of the important papers in the area (e.g., ISI Web of Knowledge (WoK) indexes all Journal Citation Report journals), with the journals of the top publishers such as IEEE, Elsevier, ACM, Springer, Taylor & Francis, and so on. In addition, the Scopus and WoK databases impose quality conditions on journals and conferences that must be met to include them, so articles found in these databases follow a minimum quality standard. Moreover, these databases are selected by accreditation agencies (e.g., the National Agency for Quality Assessment and Accreditation of Spain) because of their quality. Therefore, they can be representative enough to cover most of the top contributions in the area. Other reviews have also relied on these two databases for their analysis [62], [63].

Table 1 shows the inclusion and exclusion criteria that we used for selecting the papers for this review. First, we apply a search according to the inclusion criteria. Next, we filter the results applying the exclusion criteria to the results obtained with the inclusion criteria.

With regard to the inclusion criteria, in this case, as the area of interest is prediction in MOOCs, the search can be decomposed into two parts so that keywords have to contain the idea of "prediction" and the idea of "MOOCs." Because of that, an advanced search query was used with a global AND operator that separates both ideas and OR operators inside each one of them to include several synonyms of both ideas. The idea of prediction was covered with the terms (1) predict, (2) prediction, (3) predictive and (4) forecasting, which includes the main variants of the word "predict" (noun, verb and adjective) and a typical synonym in the literature (forecasting). Moreover, the idea of MOOCs was covered with the terms (1) MOOC, (2) MOOCs, (3) Massive Open Online Course, and (4) Massive Open Online Courses; these terms include the possibility of using the acronym or not, and the singular and plural forms. Thus, the combination of words has the following format (regardless of the database):

(predict OR prediction OR predictive OR forecasting) AND (MOOC OR MOOCs OR "Massive Open Online Course" OR "Massive Open Online Courses").

Once the query string is defined, it is important to note where to look for these words. In this case, searches were restricted to the title, abstract, and keywords, since these parts might contain the most representative terms used in the paper. Following that, 183 papers were retrieved from Scopus and 102 papers from ISI Web of Knowledge. However, 67 papers were duplicated, which gives a total of 218 different papers. Searches covered until the end of 2017 but they had no restriction on the initial date. The reason is that the MOOC phenomenon is very recent and even without a time restriction, there is an implicit "*from date*" derived from the appearance of this type of course.

Next, we explain the rest of the exclusion criteria that we followed (see Table 1). It was necessary to filter out papers that were beyond our scope. For example, there were cases where the topic was MOOCs in general and authors were predicting a rise in their popularity, but the article was not about forecasting any variable that indicates a phenomenon in the course. Furthermore, there were articles in which they only visualized existing data from the courses or they proposed learning indicators but they did not make any prediction. Therefore, such articles were discarded. A paper had to present a study in which at least a model was created to obtain a variable (outcome) from MOOC data. For example, papers that showed only correlations between variables were also excluded. Nevertheless, papers were included if at least a prediction model was built, even if the evaluation was informal (e.g., it analyzed only regression coefficients). In addition, if a paper did not contain MOOC data, it was excluded. For example, there were papers that used reviews and opinions of a MOOC outside the platform to develop models. However, they were excluded as the input was not data from the MOOC. Furthermore, studies based on online courses restricted to on-campus students were excluded even if they used typical MOOC platforms as they did not use real MOOCs. Other exclusion criteria included: position papers; secondary studies, such as surveys and systematic literature reviews; and papers not written in English.

TABLE 1
INCLUSION/EXCLUSION CRITERIA

INCLUSION CRITERIA	
	Search in titles, abstracts, and keywords in Scopus and WoK: (predict OR prediction OR predictive OR forecasting) AND (MOOC OR MOOCs OR "Massive Open Online Course" OR "Massive Open Online Courses").
EXCLUSION CRITERIA	
1	Studies published in 2018 or later
2	Duplicate papers (only one paper was included)
3	Out of scope studies
4	Studies that did not include a predictive model
5	Studies whose data set is not obtained from the MOOC
6	Studies about online courses restricted to on-campus students
7	Position papers

8	Secondary studies (e.g., surveys, systematic literature reviews)
9	Non-English papers

Following the previous criteria, three main steps were followed for each paper: 1) title and abstract were read and the article was only discarded if there was very clear evidence that the article was not related to the field of study; 2) introduction and conclusions were processed to check the validity of the different exclusion criteria (e.g., the paper is about predictive models), and 3) the part of the article was searched where the data set was described to check that data were taken from a real MOOC (and not from an online course with restricted access or a face-to-face course). When there was no clear conclusion after these steps, other parts of the article were inspected to make a decision.

There were 82 articles that met these criteria from Scopus, and 37 from ISI Web of Knowledge. In total, and after removing overlapping between the two databases, there were **88** different articles that met the aforementioned criteria (see Appendix for the details of each of the selected papers). The figure is reasonable as MOOCs are relatively recent as is the use of prediction in online courses.

3.2 Description of the set of selected papers

This section provides some general information about the 88 selected papers. The first aspect considered is whether the article was published in a journal or at a conference. Among the 88 articles, there is a clear trend for conferences since there are 72 articles released at conferences while only 16 articles are published in journals. However, the distribution of conferences and journals is very heterogeneous. In the case of journals, there is just one repetition, i.e., the 16 articles were published in 15 different journals. The *Computers in Human Behavior* was the only journal with more than one selected paper published. In the case of conferences, the 72 remaining articles were published at 39 different conferences.

Table 2 presents the number of papers that were published at each conference venue. The conferences that predominate over the rest are *Learning at Scale* (L@S) (e.g., [64], [65]) and the *International Conference on Learning Analytics & Knowledge* (LAK) (e.g., [66], [67]) with 10 articles each. As regards the rest of the conferences, only four more had at least three papers in the set (see Table 2).

TABLE 2
TOP CONFERENCES IN THE DATA SET

ID	Name	Nº
1	ACM Conference on Learning at Scale (L@S)	10
2	International Conference on Learning Analytics and Knowledge (LAK)	10
3	AAAI Conference on Artificial Intelligence (AAAI)	3
4	ACM Conference on User Modeling, Adaptation and Personalization (UMAP)	3
5	International Conference on Database Systems for Advanced Applications (DASFAA)	3
6	European MOOCs Stakeholders Summit (EMOOCs)	3

Publication dates are important to understand the popularity of prediction in MOOCs in recent years. Table 3 shows the distribution of articles over time. The distribution clearly indicates that there is an increasing interest in this field, which is particularly evident in the year 2017. While there were only six papers published in 2014, 41 papers were published in 2017. It is also worth noting that there are no articles published before 2014. This is reasonable because MOOCs are a recent phenomenon, and corroborates the fact that there was no need to specify a restriction in the initial date of the search in the databases.

TABLE 3
NUMBER OF ARTICLES PUBLISHED PER YEAR

Year	Articles
2014	6
2015	21
2016	20
2017	41

4 RESULTS AND DISCUSSION

In this section, the results for each of the research questions stated in Section 1 will be presented and discussed based on the analysis performed from the articles retrieved. Each research question is in its own subsection.

4.1 RQ1: What are the most common characteristics of the MOOCs that have been used for prediction?

The first research question is about the features of the MOOCs that appear in the articles on prediction. These features are: (1) the number of MOOCs considered in each paper; (2) the thematic areas of these MOOCs; (3) the platforms on which these MOOCs are deployed; (4) the number of enrolled users in these MOOCs; and (5) the duration of these MOOCs.

The first aspect that has been considered is how many MOOCs are analyzed in each paper. Almost half of the articles (38 out of 88) only consider one single MOOC, while the rest consider more than one MOOC to compare prediction results between several courses. Sometimes, when more than one MOOC is analyzed, it may happen that the same course is considered in several editions, while on other occasions, authors take different courses with different thematic areas to be able to evaluate the differences in results across courses. For example, Qiu et al. [68] considered 11 MOOCs from XuetangX, categorized into science and nonscience courses, with the aim of predicting certificate earners and scores. When evaluating the performance of the certificate-earning prediction with the AUC (Area Under the Curve), the value was between 94 and 95 % for both categories, considering features classified as demographic, forum activities, and learning behavior. Nevertheless, results showed a significant dif-

ference when removing behavior features, which are related to watching videos and doing assignments (the AUC value was about 92 % for science courses, and only 84 % for nonscience courses). Therefore, considering more than one course with different thematic areas can be useful for identifying patterns.

With regard to thematic areas, there was a vast list of courses in the 88 articles, and these courses were associated with many different areas. Groups of areas of knowledge that contained multiple related disciplines were used to classify the courses into thematic areas. Wu et al. [69] identified five major categories of study areas: humanities, social sciences, natural sciences, formal sciences, and the professions and applied sciences. Figure 1 shows the distribution of articles over these categories. Results show that most of the MOOCs used for prediction are related to professions and applied sciences, social sciences, and formal sciences.

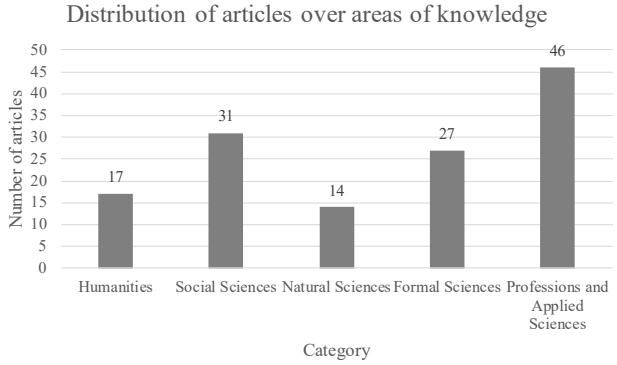

Fig. 1. Distribution of articles over areas of knowledge.

In terms of the platforms on which MOOCs are deployed, there can be many, but only six appeared at least twice in the selected papers: edX, Coursera, Open edX (including XuetangX), Canvas, Stanford platforms (e.g., Stanford Lagunita), and Telescopio. The distribution of articles over the platforms can be seen in Figure 2. In this figure, there is also an "Others" category for those papers that collect data from platforms other than the aforementioned six (e.g., MiríadaX, Iversity, etc.). In this figure, it is worth noting that Coursera and edX predominate over the rest. There are 33 articles in which MOOCs from Coursera are analyzed, and 30 in which MOOCs come from edX. This is reasonable, and it shows a correlation between the popularity of the platform [70] and the number of related papers.

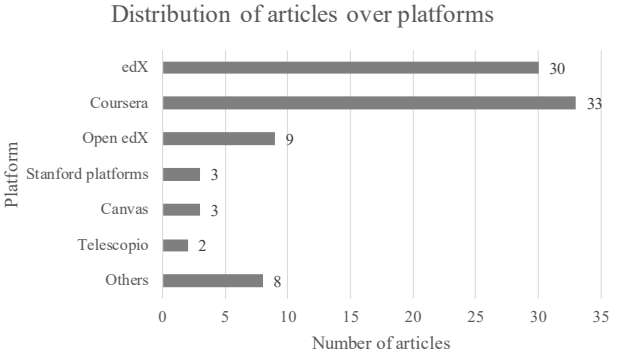

Fig. 2. Distribution of articles over the platforms.

In regard to the number of enrollees in the MOOCs analyzed, three categories were defined: small (less than 20,000), medium (between 20,000 and 60,000), and large (more than 60,000 enrollees). According to this classification and bearing in mind that one paper can be classified into more than one category (if a paper analyzes more than one MOOC and different courses fall into different categories), there are 38 articles that consider small-sized MOOCs, 30 that analyze medium-sized MOOCs, and only 11 that use data sets from large-sized MOOCs; 30 papers do not provide this information. This information is summarized in Table 4. Just to provide an example, Boyer and Veeramachaneni [71] developed a predictive model to forecast dropouts in three courses. In these cases, they considered the same MOOC offered in edX, called *Circuit and Electronics*, but in three consecutive editions with 154,753 (large), 51,394 (medium), and 29,050 (medium) learners, respectively. One of their aims was to develop transfer models and they showed that the performance in the first two courses transferred well to the third one in a better proportion than in other combinations. The number of users in each course is flagged as a possible reason (the first two courses have more learners for developing models).

TABLE 4
DISTRIBUTION OF ARTICLES ACCORDING TO THE NUMBER OF ENROLLEES

Number of enrollees	Articles
Small (under 20,000 users)	30
Medium (between 20,000 and 60,000 users)	38
Large (above 60,000 users)	11

In terms of the duration of MOOCs, Comer [72] believed that the ideal duration was between six and seven weeks. Liyanagunawardena and Williams [73] carried out a review about MOOCs on health and medicine, and the average duration was 6.7 weeks with a range of between three and 20 weeks. With this information, three categories were created for classifying the papers: short duration (less than 5 weeks), medium duration (between 5 and 8 weeks), and long duration (more than 8 weeks). Again, one paper can be classified into several categories (if it analyzes more than one MOOC). One article was classified as short, 34 as medium, and 37 as long. This is summarized in Table 5. The shortest course found was a MOOC from Curtin University hosted on edX, whose length was four weeks [74]. In that case, the article focused on predicting dropout in two editions of the course, which was relevant because more than 93 % of the enrollees did not complete the course, even though it was short. The longest MOOC was an 18-week course about electricity and magnetism, offered on edX [75]. In that case, authors developed a predictive model to forecast certification using SVMs and an adaption of LDA (Latent Dirichlet Allocation) to edX clickstream data. Results showed that LDA could be used for user modeling in

MOOCs and it was possible to achieve an accuracy of 0.81 using only the first week of logs.

TABLE 5
DISTRIBUTION OF ARTICLES ACCORDING TO THE MOOC DURATION

MOOC Duration	Articles
Short (under 5 weeks)	1
Medium (between 5 and 8 weeks)	34
Long (more than 8 weeks)	37

4.2 RQ2: What outcomes have been predicted in contributions about MOOCs?

Prediction in MOOCs can focus on several targets, such as learning outcomes and learners' behaviors. In an initial observation over the 88 selected papers, there were many categories for prediction outcomes and similar ones were grouped together. For example, predicting who is going to complete the course is almost the same as predicting who is going to drop it, so these two were grouped together. Articles that classify forum posts into different categories were grouped together as well. After this process, seven categories were considered:

- **Certificate earner.** The purpose is to predict whether a student will get the certificate at the end of the course or not. The criteria for earning a certificate can vary from one course to another. For example, some MOOCs can set the threshold at 50 % of total scoring while others might set it at 80 %. The considered dependent variables for certificate earners are categorical and normally dichotomous, although results can be grouped into discrete categorizations. For example, Xu and Yang [76] modeled students in three categories depending on their activity: certification earning (people whose aim is getting a certificate), video watching (people whose aim is acquiring knowledge), and course sampling (people who just want to have a look at the course). Afterwards, they predicted whether a student with the intention of getting a certificate would get it or not.
- **Dropout**. The aim is to predict whether a student will leave the course before completing it, so results are binary. In this group, there are also papers about course completion since it is the same concept but in a positive way. For example, Jiang and Li [77] extracted features from different sources (e.g., assignment-viewing behavior, video-viewing behavior, object-accessing behavior, etc.) and took different combinations of two views (sources) to predict dropout based on multiview ensemble learning (they trained two classifiers and predicted a dropout if the sum of the probabilities of dropping out of each one was greater than 1). They found that multiview features performed better than the aggregation of features (i.e., used a single classifier with all features together). It is worth noting that there is a difference between dropout and certificate earners. A student who completes a course might not get a certificate.
- **Scores prediction.** The purpose is to predict the score that a student will get in a certain test or course. Normally, the prediction refers to the final grade of the course but this category applies to any prediction about scores (e.g., partial scores, scores in one assessment activity, etc.). Therefore, several types of variable are possible: Variables can be continuous (if grades are expressed in a range such as 0–10), categorical (if grades are expressed from A to F), and so on. For example, Brinton and Chiang [78] developed a model with factorization machines, K-NN (K-Nearest Neighbors), and a proper algorithm to predict whether a student is Correct on First Attempt (CFA) when answering a question in the MOOC. In this case, students can answer the question several times but the variable is binary (a question can be right at the first attempt or not).
- **Forum posts classification.** The aim is to predict (or classify) the category of a post in the forum according to different classifications, so dependent variables are normally categorical and nominal. For example, Arguello and Shaffer [79] took a data set from a course on *Metadata* from Coursera and classified posts into seven categories according to their speech act: question, answer, issue, issue resolution, positive acknowledgement, negative acknowledgment, and other.
- **Relevance of content.** The purpose is to identify identify whether the elements presented in the course (videos, lectures…) are relevant or of interest for the learners. For instance, Yang, Adamson, and Rosé [80] proposed a context-aware matrix factorization model to predict learners' preferences over the forum questions. In this case, dependent variables can be either continuous or categorical.
- **Student behavior.** The purpose is to identify characteristics of learners related to their behavior in the course. In this case, dependent variables are usually categorical. As there can be many possible behaviors, two examples are given. For instance, Hicks et al. [81] classified learners into five different categories (fully engaged, consistent, two-week, one-week, and sporadic) according to their engagement based on data from platform use and a pre-survey. Moreover, Bote-Lorenzo and Gómez-Sanchez [82] developed different models to predict whether the value of three engagement indicators (video, exercise, and assignment) decreased at the end of a chapter with respect to the previous one or not.
- **Others.** This category applies to any purpose that has not been identified in the previous six categories. For example, Vu, Pattison, and Robins [83] proposed a model to predict whether a student will have future post events in the forum. Some of the results show that people who ask for information in the forum and users with a higher degree in the network (which is related to the number of threads to which a learner has posted) are more likely to post again.

These seven categories enable the identification of trends on the most popular outcomes in MOOC prediction. Figure 3 shows the distribution of these outcomes in the set of papers considered in this review. The most

common category is dropout. This is reasonable as although this problem also occurs in more formal contexts and has been dealt with since the 1960s [84], nonformal MOOC contexts are more susceptible to dropouts because of the diversity of participants, their motivations, and the lack of teacher support.

The second and third categories are scores prediction and certificate earners (both with a similar number of articles), which means that there is a high interest in knowing learning outcomes, either the specific score or whether the learner has passed the course (and got the certificate) or not. Then, there is interest in predicting specific student behaviors and inferring some data by classifying forum posts. Finally, there are a few papers that assess the relevance of the educational materials, and the category that covers other outcomes.

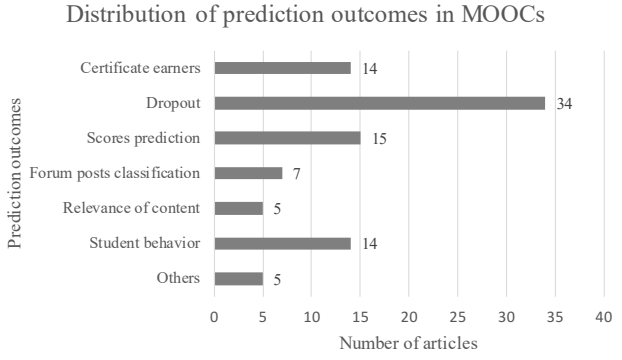

Fig. 3. Distribution of prediction outcomes in MOOCs.

4.3 RQ3: What are the prediction features that are used to build prediction models in MOOCs?

Prediction features serve to build models aimed at predicting the aforementioned outcomes. Prediction features can consider several elements of a MOOC, and as each prediction model can use several features, the full list of them can be quite long. For this reason, prediction features have been grouped into seven possible categories of variables:

- **Demographic variables.** These are related to general characteristics of the learner (e.g., age, level of schooling, country of origin, primary language, employment status, etc.). For example, Al-Shabandar et al. [85] developed a predictive model to forecast course outcome using different features, and including participants' demographic information, such as age, gender, and educational background. This demographic information can be taken directly from the platform, but sometimes is obtained from a survey completed by the learners. For instance, Greene, Oswald, and Pomerantz [86] used a pre-course survey to collect some data for predicting dropouts. Their demographic information included age, level of schooling, academic experience in the subject area of the MOOC, and familiarity with course topic. It is important to note that surveys usually collect some data about learners' commitment or expectations that are not demographic and should be excluded from this category. Thus, data from a survey are only eligible for this category if they include real demographic information for the study.
- **Video-related variables.** These are related to logs generated about what learners do when interacting with videos on the platform. Brinton et al. [87] considered logs from different video events, such as play, pause, skip back, skip forward, change to faster rate, change to slower rate, and change to default rate, to predict CFA scores. Other traditional variables include: video lecture downloads, number of lecture views, and total time or percentage of video completed over the course.
- **Exercise-related variables.** These are related to logs generated about what learners do when interacting with the exercises (e.g., quizzes, questionnaires, exams) on the platform. Boyer et al. [88] identified several variables of this category, which include: number of distinct problems attempted, number of submissions (in this case, a submission corresponds to a problem attempt), number of distinct correct problems, average number of submissions per problem, ratio of total time spent to number of distinct correct problems, ratio of number of problems attempted to number of distinct correct problems, and average time between a problem submission and a problem due date, among others.
- **Forum-related variables.** These variables are related to the forum activity. There are hundreds of them, and as the text can be processed to obtain new variables as well as social relationships between users, forums are a source from which much information can be extracted. Klüsener and Fortenbacher [89] only used forum-related variables to predict whether a student was going to be successful or not in a course; some of the features they used were: number of negative ratings, number of positive ratings, number of hyperlinks, number of images, number of words, number of received positive ratings, number of received negative ratings, number of answers, number of comments, and number of posts. Arguello and Shaffer [79] presented a long list of features that could be extracted from posts, including: measures of positive and negative sentiments, measures concerning whether the author is expressing uncertainty or is comparing something, linguistic properties such as number of words, words per sentence, frequency counts for pronouns, measures of nonfluencies and fillers (e.g., "er," "blah"), text similarity features where TF-IDF was considered, temporal features, and social features, among others. In the last area, several features can be used, such as the degree, or measures related to centrality (e.g., betweenness, closeness) or prestige (e.g., authority, hub).
- **Platform use variables.** These are related to general variables about what actions a user takes on the platform, or when he/she takes such actions, but not specifically related to videos, forum posts, or activities. In this category, Liang et al. [90] included variables such as: time elapsed after first activity, number of accesses, number of periods when the user has been on the course, number of weeks the student spends on the course, last access time, registration time, total click

count, course access interval, access interval for categories (video, exercises…), total engagement time, or average engagement time per session.

- **Survey variables.** These are features obtained from a survey, excluding demographic information. They are normally about users' interests, commitment to finishing the course or obtaining a certificate, or expectations about the course. For example, Zhong et al. [91] predicted learners' styles according to three categories: active, passive, and both active and passive. In order to do that, they used questionnaire survey data and activity data. The survey included some question related to the attitudes and feelings of students toward learning on a MOOC (e.g., "What are the advantages of learning in MOOCs?").
- **Other variables.** These contain the variables that cannot be categorized in the previous groups. For example, Hong, Wei, and Yang [92] predicted dropouts in 39 courses from XuetangX and used features such as the rate of users' dropout (i.e., number, rate, and scores of courses from which users dropped out) and the rate of class dropout (i.e., number, rate, and scores of dropout users). Moreover, Robinson et al. [42] took data from open-ended responses and combined all text to form a single document for each student. Then, they used Natural Language Processing to obtain a matrix where each student was represented by a row and columns contained words or phrases with the number of appearances. In this case, this is similar to some forum-related variables, but as they are from another source, they have been considered in the category "Other variables."

Figure 4 shows the distribution of prediction features in the MOOCs considered in the set of papers. Papers may fall into several categories if they use several types of variables. Results show that there are four categories that stand out (platform use, forum-related, exercise-related, and video-related), with platform use being the top category. This demonstrates the importance of knowing when users access the platform or the clicks they do to build prediction models. However, results also show that the general trend is to use a combination of different types of features in order to gain prediction power, so different sources might be taken whenever possible. An important finding is that demographic variables (and all those obtained from a survey) are not as important as the rest of the features: Their predictive power is generally lower than that of activity data obtained from the tracking logs of the platform, as was demonstrated by Brooks, Thompson, and Teasly [93].

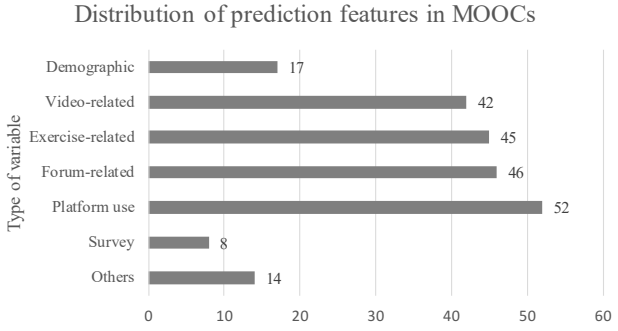

Fig. 4. Distribution of prediction features in MOOCs.

4.4 RQ4: What are the techniques/models used for prediction in MOOCs?

Prediction features can be used to obtain the outcome that is being forecast, but to do so, a subjacent model that relates both prediction features and outcomes is needed. In a machine learning analysis, feature selection and extraction is considered relevant, but the technique used and how parameters are tuned for training the algorithm are also important. The question here refers to the models used for predicting in MOOCs.

The list of techniques used in the papers is very long. This means that there can be many options for implementing the algorithms for prediction. Moreover, sometimes several models are used in the same paper to compare the results. In this case, the clustering criteria have been to group related algorithms and consider a special group, *Others*, for algorithms that appear only once or twice in the entire set of articles. In this case there are eight categories:

- **Regression.** Makhabel [94] defined the aim of regression as finding the best curve in a multidimensional space that fits sample data. Regression is one of the simplest and most commonly used methods for prediction, and is usually taken as a baseline. For example, Wise et al. [95] used L2 regularized logistic regression to classify threads depending on whether these were substantially related to course content or not.
- **Support Vector Machines (SVMs)** [96]. These are a very common supervised learning technique. Their aim is to find a hyperplane that maximizes the margin between classes to create classifications. In the context of prediction in MOOCs, they have been used for several purposes. For example, Fei and Yeung [97] used SVMs for predicting dropouts in two courses from Coursera and edX. Macina et al. [98] used SVMs to predict students' willingness to answer a question.
- **Decision trees (DTs)** [99]. These are methods for supervised learning whose goal of prediction is performed by learning simple decision rules from the prediction features. Among the papers set, Gardner and Brooks [100] used DTs to predict whether a learner would register any activity in the third week of the course. These authors used both multilayer perceptron and DTs as models, and compared results with different metrics. Results showed that not all metrics reflected a true difference in model performance; the accuracy in both models was 0.83, while the AUC was between 0.57 (with DTs) and 0.60 (with multilayer perceptron).
- **Random Forest (RF).** This is an ensemble of unpruned trees created by bootstrapping samples of the training set and performing random feature selection in tree induction. Final predictions are made by aggregating

the predictions of the ensemble [101]. An example of its use in MOOCs is the work Laveti et al. [102] did to forecast dropouts. These authors developed different single models, and RF offered the best performance with an AUC of 0.875. Nevertheless, they managed to improve the AUC up to 0.91 when using a stacked ensemble using all the algorithms, although that ensemble produced more computational overhead, which was a problem for scalability in MOOCs. Moreover, Ye et al. [64] compared RF with other algorithms, such as logistic regression, SVMs, and DTs to forecast dropouts, and found that RF consistently performed better than the others.

- **Naive Bayes.** This is a Bayesian classifier (based on the application of the Bayes' theorem), which assigns samples with the most likely class according to the features, assuming that features are independent given the class [103]. As an example, Lu et al. [104] used several predictive models, including Naive Bayes, to forecast whether a learner was going to take the final exam of the MOOC or not. The study used three courses and the best model differed for each case, although Naive Bayes achieved the best recall in one of them.
- **Gradient Boosting Machine (GBM).** This is a machine learning technique, which generates a strong predictive model using an ensemble of weak predictive models [105]. Typical uses are with DTs and regression. For example, Ruipérez-Valiente et al. [106] used different models for predicting certificate earners in an edX course. One of their aims was to analyze how the predictive power evolved over weeks and they found GBM to be the most stable and best or second best in terms of performance over the first four weeks, which was considered the most important period in the course (as predicting later would not be effective in supporting learners).
- **Neural Networks.** These are parallel computing systems for optimization and learning based on the structure and functionality of the brain [107]. They have recently gained popularity in the area of prediction in MOOCs (12 out of the 14 papers using them were published in 2017). As an example, Pérez-Lemonche, Martínez-Muñoz, and Pulido-Cañabate [108] predicted grades in a MOOC about Android programming on edX using RF and Neural Networks. They found that it was possible to predict the score with both algorithms with around 10 % of mean absolute error.
- **Others.** This category applies to every model that is not included in the previous categories. These models include General Bayesian Networks, factorization machines, K-NN, PSL (Probabilistic Soft Logic), K-means and Gaussian Processes, among others. For example, Wang et al. [109] adapted the Bayesian Knowledge Tracing (BKT) model to include Knowledge Components (KCs) and developed their novel methods Multi-Grained-BKT and Historical-BKT. In another article, Arabshahi et al. [110] developed a latent tree structure over KCs and predicted student learning. For doing that, they introduced a parametric model class, the Conditional Latent Tree Model (CLTM), which takes into account different factors for predicting high-dimensional time series.

The distribution of prediction techniques used in the set of papers can be seen in Figure 5. Results show a clear evidence of the heterogeneity of the different models as there are many articles in the *Others* group. Nevertheless, the category that predominates is regression. The possible reasons for that are that many articles use this model because of its simplicity, and that some others use it as a baseline to compare the performance with other techniques. Apart from that, many articles that use SVMs and RF have been identified. This may suggest that they may be a good option to try when exploring different models to predict a variable in a MOOC. Finally, it is worth noting that many papers about neural networks appeared in 2017. This suggests that they can be of interest in the area of prediction in MOOCs, and the use of deep learning may even increase their use in the future.

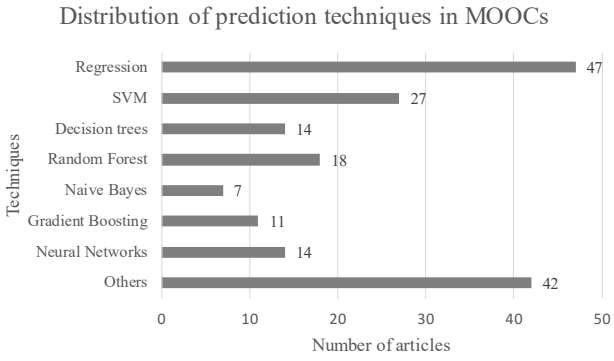

Fig. 5. Distribution of prediction techniques in MOOCs.

4.5 RQ5: What metrics have been used to evaluate prediction results in MOOCs?

After training the model and making predictions about different outcomes, there is a last step that needs to be considered: the evaluation of results. This is important to check that models generalize well and prediction results are reliable. There can be many different metrics but seven of them predominate in the set of articles that has been analyzed. Figure 6 shows the distribution of metrics that have been used to evaluate prediction results in MOOCs.

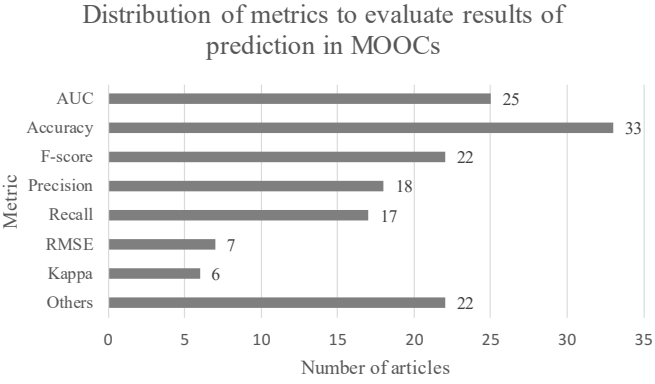

Fig. 6. Distribution of metrics to evaluate results of prediction in

MOOCs.

The most common metrics are: accuracy and AUC (Area under the Curve). A ROC (Receive Operating Characteristic) curve shows the relationship between true positive rates and false positive rates with various threshold settings. The quality of the predictor can be measured with the AUC. A possible classification of the AUC results appears in Table 6 [111].

In the papers considered, 25 used the AUC to measure the performance but there were very heterogeneous results. There were papers that reported results ranging from very poor to excellent, and different results could even be obtained in the same paper when considering different experiments or courses. For example, Xing et al. [34] predicted dropouts in the different weeks of an eight-week course hosted by Canvas. Although they got an AUC of 0.807 in the first week, when they forecasted in the seventh week (which is the last opportunity for making worthwhile predictions) they obtained an AUC of 0.961. Moreover, Qiu et al. [68] also obtained an excellent AUC when they got about 0.95 in grades prediction for science courses.

TABLE 6
ASSESSMENT CRITERIA FOR AUC RESULTS

ROC curve AUC	Quality
0.9 < AUC ≤ 1.0	Excellent
0.8 < AUC ≤ 0.9	Good
0.7 < AUC ≤ 0.8	Fair
0.6 < AUC ≤ 0.7	Poor
0.0 ≤ AUC ≤ 0.6	Fail

Accuracy, which is the proportion between true cases (positives and negatives) and the total number of cases considered, has also appeared in many research works. Results have been variable, and for example, Brooks, Thompson, and Teasly [112] achieved an accuracy of 0.91 in a MOOC when predicting whether a learner was going to achieve a distinction in the course (grade above 85 %). However, when they used the same model in a second edition of the same course, the accuracy dropped to 0.65.

Two other popular metrics are recall and precision. Recall measures the fraction between true positives and false negatives while precision measures the relationship between true positives and false positives. In both cases, results have been very heterogeneous. One reason for this is that researchers have combined MOOCs and features, which leads to a variety of possible results.

The combination of recall and precision provides another well-known metric: F-score. In the set of papers, F-score has been used more times than recall and precision separately, F-score allows results to be compared combining both values. In some cases, the three values (recall, precision, and F-score) are presented together. For example, Ramesh et al. [45] classified forum posts according to different course aspects, which included videos, quizzes, social interactions, and the certificate, and a fine classification of each aspect (e.g., video messages were classified into those related to the video quality, audio, subtitles, etc.). They also classified messages according to their sentiments, which could be positive, negative, or neutral, and reported all their results in terms of recall, precision, and F-score. This case is also a good example to show the variety of results because although they obtained an F-score of 0.706 in course aspects related to quizzes, they only achieved 0.189 in positive sentiments.

Another metric that has been used in some articles is the RMSE (Root-mean-square error), which represents the sample standard deviation of the difference between predicted and real values. For example, Elbadrawy et al. [113] used RMSE for performance evaluation of the predicted grades of homework activities in a MOOC, considering previous graded activities attempted by the learner. However, one possible problem with this measure is that its value can be very different depending on the samples, and the same RMSE value can have multiple interpretations depending on the data (i.e., RMSE=1 when grades are from 0 to 10 is not the same as when grades are from 0 to 100). For this reason, it is very important to specify the ranges of the data. Brinton et al. [59] predicted CFA (Correct on First Attempt) when answering questions and they clearly specified that each sample could have a value of 0 or 1, so their 0.438 RMSE value obtained using all users over the full course can be interpreted without ambiguity.

In some cases, Cohen's kappa has also been used for reporting results. This metric calculates pairwise agreement among a set of coders (i.e., humans or machines that assign labels to data) making category judgements [114]. It takes into account a correction for expected change agreement. Its range can vary between -1 and 1 and Table 7 indicates the level of agreement depending on the value, according to the categories proposed by Landis and Koch [115]. An example of its use was the article developed by Chaplot, Rhim, and Kim [116]. They predicted student attrition in MOOCs using several features, including the student sentiment obtained through a sentiment analysis algorithm with forum posts. They managed to obtain a kappa of 0.432, which was moderate, although significantly better than other previous contributions.

TABLE 7
ASSESSMENT CRITERIA FOR COHEN'S KAPPA RESULTS

Kappa	Agreement
0.81–1.00	Almost perfect
0.61–0.80	Substantial
0.41–0.60	Moderate
0.21–0.40	Fair
0.00–0.20	Slight
Lower than 0	Poor

Finally, there is the category *Others*, which includes other metrics different to the previous ones, although there can be similarities. For example, Crossley et al. [66] used the confusion matrix to present results about course completion. This matrix provides the true and false positives and negatives, so other previous metrics can be obtained from it. Another approach was measuring the R^2 to

measure the proportion of the variance that is explained by the model. For instance, Kennedy et al. [41] reported that their model to predict scores achieved an R^2 of approximately 91 %.

To summarize, there are different approaches to evaluating the performance of prediction models in MOOCs, but the overall idea is to explore the accuracy, the AUC in the ROC curve, the kappa, or variables that use true and false positives or negatives in several ways (such as recall, precision, or F-score). Among these variables, accuracy clearly predominates over the rest. Nevertheless, it is important to note that this metric might not be appropriate in many cases. For example, if the dropout rate is 95 %, the accuracy could be 0.95 (which could seem to be good) by just predicting the most likely class; the metric, however, would be inappropriate in this case. To guide the metric selection, a researcher could refer to the work by Pélanek [117], who suggests the most appropriate metric for each situation.

5 Future Directions

The analysis of the set of selected papers in relation to the research questions served to provide an overview of what has been done in prediction in MOOCs. This can be useful for identifying areas where more research is required. This section provides possible future directions to support researchers in their work, and these are presented together with the main findings. We classify futures lines of work into three groups depending on their priority: high, medium, and low. The reason for this is that the lack of existing research in an area does not necessarily mean that there is a need for more work in that area. Therefore, priorities could be established among aspects in which there is a lack of research. This way, this section can suggest where the field of prediction in MOOCs maybe headed. In particular, our view is that predictive power has a lot of room for improvement and that improvements in this direction can be significant since it is very important to have accurate predictions. In addition, we consider very important the prediction of new interesting outcomes and the relationships among them since this would allow different aspects of the learn-ing process to be anticipated and improved. The list of future research lines, in order of priority, is as follows:

HIGH PRIORITY

1. **New features could be incorporated in predic-tive models.** The study revealed that there are many different low-level indicators that are used as prediction features. Among these prediction features, clickstream data about platform use seem to be the most interesting feature for researchers as many contributions used them, while forums can also be a good place to look for new indicators, as text can be analyzed in many ways. Moreover, this analysis revealed that although there are many low-level indicators (from the same set of categories) used in published papers as prediction features, there is a need to include other higher-level indicators as prediction features such as self-regulated learning variables on MOOCs, such as those presented by Kizilcec, Pérez-Sanagustín, and Maldonado [118]. The inclusion of these new features in the models could improve the prediction power. Furthermore, results suggest that pre-survey data could be included as predictors in predictive models, although these variables should be analyzed carefully because they could be biased.
2. **New machine learning techniques and predictive models can enhance the predictive power.** Results showed that there are no clear predominant prediction techniques to be used for education. Many studies have considered regression models because of their simplicity, as they can be good as a baseline. SVMs and RF also seem to be used widely. In any case, there is a high heterogeneity in the prediction techniques. One interesting finding is the increase of popularity in neural networks, which means that deep learning could be expanded along with SVMs and ensemble and boosting methods, perhaps achieving higher predictive powers. As a future research direction, new prediction techniques can be used for different purposes, which can combine different models (e.g., using ensemble methods). Contributions might focus on a combination of developing new models and enhancing the predictive power of current ones through improvements (e.g., by adding new features). Furthermore, research may also analyze when is the best mo-ment to predict since achieving perfect predictions at the last moment could be too late to have an effect on learners and their behavior.
3. **More prediction outcomes might be consid-ered and the relationship between outcomes could be analyzed.** The results of the study showed that dropout is the most important prediction outcome. Currently, attrition rates in MOOCs are in most cases over 87 % [119], these being much higher than in closed courses. This raises interest in the early detection of student dropouts in order to perform interventions that may reduce them. Nevertheless, it is worth noting that dropouts in MOOCs are not always a problem since these courses are also open to explorers. Additionally, the analysis showed a high interest in learners' results, which can also be used to detect where learners face difficulties with a view to improving the contents of the MOOC or making early interventions. However, there is a lack of contributions concerning prediction of the learning efficiency (i.e., whether a student learns at a good pace), learners' behaviors (e.g., persistence or constancy), or learners' expectations. Furthermore, it would also be interesting to predict what parts of the course are more challenging for students and what parts are unnecessary. What is more, most of the papers are focused on the prediction of just a single variable and there is also a lack of studies that predict more than one variable and establish relationships among them, and this could be addressed in future research.

MEDIUM PRIORITY

4. **Predictive models can be more generalizable.** The study showed that almost half of the articles tested predictive models in only one MOOC. This is a limitation of the results since it is not possible to determine how generalizable the predictive models are. A future research line could focus on designing adaptive models that can work in different courses, with different students, platforms, etc. This line of work will require a better understanding of each context to identify how to adapt models easily to be transferable. Moreover, making models generalizable would also facilitate the enhancement of the predictive power, since models would be more stable across different courses.
5. **Evaluation of predictive models can be improved.** Results showed that performance evaluation typically uses related metrics**.** Accuracy is the most adopted metric for performance evaluation of prediction models in MOOCs, followed by AUC, although common metrics calculated from the confusion matrix (e.g., recall, precision) [120] are also extended. RMSE is also used when predicting continuous variables, although there are more cases of classification problems, either binary or multiclass (e.g., grades on a scale A to F). Nevertheless, accuracy can sometimes be inappropriate, as suggested by Pelánek [117], and it is important to carry out further research on what metrics should be used in each case, so that the results presented in the literature can be better understood. In this line, it would also be interesting to analyze whether it would be possible to use common metrics, or at least use some of them as a standard, so all papers include them for consistency.

LOW PRIORITY

6. **More research is needed in the areas of natural sciences and humanities.** There is a clear trend to analyze courses about professions and applied sciences, followed by social sciences and formal sciences. However, more efforts might be made in other areas of knowledge as it has been shown that prediction results can be very different depending on the specific thematic area. Moreover, research could focus on com-paring MOOCs from different thematic areas and cre-ating models that are transferable among courses.
7. **More research is needed in short courses.** Only one paper contained data from a MOOC whose duration was under five weeks (short course). It would be interesting to explore whether learners' behaviors change when they need less effort to complete the course. Moreover, the detection of dropouts could be studied in a different way since the margin for detecting learners at risk is lower as the end of the course is near the beginning.
8. **More research could be done with platforms other than edX and Coursera.** The analysis has shown that most of the contributions came from Coursera and edX, the two most popular platforms. This reflects the current trends in platform use and suggests the need for research on other platforms. Moreover, interoperability solutions might be designed in the future to facilitate the transfer of predictive models from one platform to another.

In addition to what has been identified from the re-search questions, it is worth highlighting that research in this area can help us to better understand learners' attitudes, preferences, and behaviors. This can benefit not only learners, but also the academic community involved in the development of MOOCs. Nevertheless, research on prediction in MOOCs could be interdisciplinary to be effective, and different stakeholders could be involved, including instructors, data scientists, educators, psychologists, etc. For example, in the case of forum data, there has been a long tradition of research coming from linguistics and educational research, based on pedagogical theories and educational models (e.g., educational experience [121]). However, these models could be com-bined with text-mining techniques and an interdiscipli-nary and convergent analysis is needed to search for reli-ability and validity. Additionally, stakeholders should note that data on MOOCs can be noisy because of the high number of learners who enroll but are not active. Because of that, it is important to clean data and focus on active learners.

Furthermore, ethical issues could arise when taking data from a MOOC, and researchers should be aware of protocols and local/national laws in order to meet regulations while doing their research. In addition, the dimensions to address on prediction in MOOCs should be clear. In the future, researchers could focus on more personalized solutions to develop models and alert learners individually. Contextualization would also be important and there could be a trade-off between its increase and the search for generalizable solutions. Finally, contributions might focus on how to make interventions on time, as interventions are the key element to make models have an impact at both institutional and educational levels.

6 Conclusions

This study has analyzed the most important articles related to prediction in MOOCs. The methodology applied allowed the main existing works to be retrieved and showed that this field is emerging as there has been a clear evolution in the number of publications in recent years. In the future, with the expansion of online education and learning analytics, which is now in its early stages (awareness and experimentation according to the model presented by Siemens, Dawson, and Lynch, [122]), more work is expected to appear and the interest in this topic will increase even more.

As well as the limitations related to the current state and adoption of prediction in MOOCs, it is important to mention that in this paper, courses restricted to students on campus with a MOOC-like format were excluded. They may share certain similarities with MOOCs and it would be interesting to focus on them as well, but the characteristics of learners are different. Moreover, although we have used two high-quality and comprehen-

sive databases, as in any systematic review it is not possible to reach 100 % coverage. Nevertheless, the authors believe that the results are representative and reflect the current state of the art.

Finally, it is important to note that based on the conclusions obtained, there is a lot of room for research in this area. This review has opened the way to focus on the areas where there can be more opportunities to innovate. We believe that the advancement of predictive models would allow the combination of these models with explanatory models to support the final step after predictions: interventions. Researchers should not neglect the intervention step as it will serve to obtain feedback in conjunction with predictions to improve learning processes.

ACKNOWLEDGMENT

This work has been co-funded by the Erasmus+ Programme of the European Union, projects MOOC Maker (561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP), SHEILA (562080-EPP-1-2015-BE-EPPKA3-PI-FORWARD), LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP) and COMPETEN-SEA (574212-EPP-1-2016-1-NL-EPPKA2-CBHE-JP), by the Madrid Regional Government, through the eMadrid Excellence Network (S2013/ICE-2715), and by the Spanish Ministry of Economy and Competitiveness, projects RESET(TIN2014-53199-C3-1-R) and Smartlet (TIN2017-85179-C3-1-R). The latter is financed by the State Research Agency in Spain (AEI) and the European Regional Development Fund (FEDER). It has also been supported by the Spanish Ministry of Education, Culture and Sport, under a FPU fellowship (FPU016/00526).

REFERENCES

[1] A.M. Kaplan and M. Haenlein, "Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster", *Business Horizons*, vol. 59, no. 4, pp. 441-450, July-August 2016.

[2] G. Christensen, A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E.J. Emanuel, "The MOOC phenomenon: Who takes massive open online courses and why?", *SSRN,* https://ssrn.com/abstract=2350964. 2013.

[3] D. Coetzee, A. Fox, M.A. Hearst, and B. Hartmann, "Should your MOOC forum use a reputation system?", *Proc. ACM conference on Computer supported cooperative work & social computing (CSCW '14)*, pp. 1176-1187, Feb. 2014, doi: 10.1145/2531602.2531657.

[4] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, "Attrition in MOOC: Lessons learned from drop-out students", *International Workshop on Learning Technology for Education in Cloud (LTEC'14)*, pp. 37-48, Sep. 2014, doi: 10.1007/978-3-319-10671-7_4.

[5] C.P. Rosé, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer, "Social factors that contribute to attrition in MOOCs", *Proc. ACM conference on Learning@ scale (L@S '14)*, pp. 197-198, Mar. 2014. doi: 10.1145/2556325.2567879.

[6] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor, "Learning latent engagement patterns of students in online courses", *Proc. AAAI Conference on Artificial Intelligence (AAAI '14)*, pp. 1272-1278, 2014.

[7] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting Instructor's Intervention in MOOC forums", *Proc. Annual Meeting of the Association for Computational Linguistics (ACL '14)*, pp. 1501-1511, 2014.

[8] R.F. Kizilcec and E. Schneider, "Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale", *ACM Transactions on Computer-Human Interaction*, vol. 22, issue 2, article 6, Apr. 2015.

[9] A.M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques", *Procedia Computer Science*, vol. 72, pp. 414-422, Dec. 2015.

[10] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-Term Student Performance Prediction: A Recommender Systems Approach", *arXiv preprint arXiv:1604.01840*. 2016.

[11] 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", *The Internet and Higher Education*, vol. 28, pp. 68-84, Jan. 2016.

[12] E. Aguiar, G.A.A. Ambrose, N.V. Chawla, V. Goodrich, and J. Brockman, "Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence", *Journal of Learning Analytics*, vol. 1, no. 3, pp. 7-33, 2014.

[13] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", *Computers & Education*, vol. 53, issue 3, pp. 950-965, Nov. 2009.

[14] A. Daud, N.R. Aljohani, R.A. Abbasi, M.D. Lytras, F. Abbas, and J.S. Alowibdi, "Predicting Student Performance using Advanced Learning Analytics", *Proc. International Conference on World Wide Web Companion (WWW'17 Companion)*, pp 415-421, Apr. 2017.

[15] Y. Neumann, E. Neumann, and S. Lewis, "The Robus Learning Model with a Spiral Curriculum: Implications for The Educational Effectiveness Of Online Master Degree Programs", *Contemporary Issues in Educational Research*, vol. 10, no. 2, pp. 95-108, 2017.

[16] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting grades", *IEEE Trans. on Signal Processing*, vol. 64, no. 4, pp. 959-972, Feb.15, 2016, doi: 10.1109/TSP.2015.2496278.

[17] S. Huang, and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models", *Computers & Education*, vol. 61, pp. 133-145, Feb. 2013.

[18] M.S. Sajjadi, M. Alamgir, and U. von Luxburg, "Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines", *Proc. ACM Conference on Learning@ Scale (L@S '16)*, pp. 369-378, Apr. 2016, doi: 10.1145/2876034.2876036.

[19] A. Pardo, N. Mirriahi, R. Martinez-Maldonado, J. Jovanovic, S. Dawson, and D. Gašević, "Generating actionable predictive models of academic performance", *Proc. International Conference on Learning Analytics & Knowledge (LAK '16)*, pp. 474-478, Apr. 2016, doi: 10.1145/2883851.2883870.

[20] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "Neural Network Approach for Students' Performance Prediction", *Proc. International Conference on Learning Analytics & Knowledge (LAK '17)*, pp. 598-599, Mar. 2017.

[21] F. Marbouti, H.A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading", *Computers & Education*, vol. 103, pp. 1-15, Dec.

2016.

[22] M.G. Brown, R.M. DeMonbrun, S. Lonn, S.J. Aguilar, and S.D. Teasley, "What and when: the role of course type and timing in students' academic performance", *Proc. International Conference on Learning Analytics & Knowledge (LAK '16)*, pp. 459-468, Apr. 2016, doi: 10.1145/2883851.2883907.

[23] C. Romero, M.I. López, J.M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums", *Computers & Education*, vol. 68, pp. 458-472, Oct. 2013.

[24] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining social media data for understanding students' learning experiences", *IEEE Transactions on Learning Technologies*, vol. 7, issue 3, pp. 246-259, Jul/Sep 2014, doi: 10.1109/TLT.2013.2296520.

[25] M. Xenos, "Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks", *Computers & Education*, vol. 43, issue 4, pp. 345-359, Dec. 2004.

[26] S. Kolowich, "Coursera takes a nuanced view of MOOC dropout rates", *The chronicle of higher education*, http://www.chronicle.com/blogs/wiredcampus/coursera-takes-a-nuanced-view-of-mooc-dropout-rates/43341. 2013.

[27] R.J. Rosen, "Overblown-claims-of-failure watch: How not to gauge the success of online courses", *The Atlantic*, https://www.theatlantic.com/technology/archive/2012/07/overblown-claims-of-failure-watch-how-not-to-gauge-the-success-of-online-courses/260159. 2012.

[28] R.F. Kizilcec, J.N. Bailenson, and C.J. Gomez, "The Instructor's face in video instruction: Evidence from two large-scale field studies", *Journal of Educational Psychology*, vol. 107, no. 3, pp. 724-739, Aug. 2015.

[29] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in MOOCs using learner activity features", *Experiences and best practices in and around MOOCs*, vol. 7, pp. 3-12, Mar. 2014.

[30] C. Taylor, K. Veeramachaneni, and U.M. O'Reilly, "Likely to stop? Predicting stopout in massive open online courses", *arXiv preprint arXiv:1408.3382*. 2014.

[31] R. Kizilcec and S. Halawa, "Attrition and achievement gaps in online learning", *Proc. ACM Conference on Learning@ Scale (L@S '15)*, pp. 57-66, Mar. 2015.

[32] J.L. Santos, J. Klerkx, E. Duval, D. Gago, and L. Rodríguez, "Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses", *Proc. International Conference on Learning Analytics and Knowledge (LAK '14)*, pp. 98-102, Mar. 2014, doi: 10.1145/2567574.2567627.

[33] T. Zhang, and B. Yuan, "Visualizing MOOC User Behaviors: A Case Study on XuetangX", *Proc. International Conference on Intelligent Data Engineering and Automated Learning (IDEAL '16)*, pp. 89-98, Oct. 2016, doi: 10.1007/978-3-319-46257-8_10.

[34] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization", *Computers in Human Behavior*, vol. 58, pp. 119-129, May 2016.

[35] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods", *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP '2014)*, pp. 60-65, Oct. 2014, doi: 10.3115/v1/w14-4111.

[36] R. Cobos and V. Macías Palla, "edX-MAS: Model Analyzer System", *Proc. International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '17)*, Oct. 2017.

[37] C. Ye and G. Biswas, "Early Prediction of Student Dropout and Performance in MOOCSs Using Higher Granularity Temporal Information", *Journal of Learning Analytics*, vol. 1, no. 3, pp. 169-172, 2014.

[38] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models", *arXiv preprint arXiv:1605.02269*, 2016.

[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 Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 716-728, Aug. 2017.

[40] T. Sinha, and J. Cassell, "Connecting the Dots: Predicting Student Grade Sequences from Bursty MOOC Interactions over Time", *Proc. ACM Conference on Learning@ Scale (L@S '15)*, pp. 249-252, Mar. 2015, doi: 10.1145/2724660.2728669.

[41] G. Kennedy, C. Coffrin, P. De Barba, and L. Corrin, "Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance", *Proc. International Conference on Learning Analytics and Knowledge (LAK '15)*, pp. 136-140, Mar. 2015, doi: 10.1145/2723576.2723593.

[42] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting student achievement in MOOCs with natural language processing", *Proc. International Conference on Learning Analytics & Knowledge (LAK '16)*, pp. 383-387, Apr. 25, doi: 10.1145/2883851.2883932.

[43] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, "DropoutSeer: Visualizing Learning Patterns in Massive Open Online Courses for Dropout Reasoning and Prediction", *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST '16)*, pp. 111-120, Oct. 2016.

[44] 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 transactions on Learning Technologies*, vol. 7, issue 4, pp. 346-359, Oct/Dec 2014, doi: 10.1109/TLT.2014.2337900.

[45] A. Ramesh, S.H. Kumar, J.R. Foulds, and L. Getoor, "Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums", *Proc. Annual Meeting of the Association for Computational Linguistics (ACL '15)*, pp. 74-83, Jul. 2015, doi: 10.3115/v1/p15-1008.

[46] A. Bakharia, "Towards Cross-domain MOOC Forum Post Classification", Proc. ACM Conference on Learning@ Scale (L@S '16), pp. 253-256, Apr. 2016, doi: 10.1145/2876034.2893427.

[47] C. Alario-Hoyos, P.J Muñoz-Merino, M. Pérez-Sanagustín, C. Delgado Kloos, and H.A. Parada G, "Who are the top contributors in a MOOC? Relating participants' performance and contributions", *Journal of Computer Assisted Learning*, vol. 32, issue 3, pp. 232-243, Jun. 2016.

[48] E. Er, M.L. Bote-Lorenzo, E. Gómez-Sánchez, Y. Dimitriadis, and J.I. Asensio-Pérez, "Predicting Student Participation in Peer Reviews in MOOCs"; *Proc. European MOOCs Stakeholders Summit (EMOOCS '17)*, pp. 65-70, May 2017.

[49] M.R. Barrick and M.K. Mount, "The Big Five Personality Dimensions and Job Performance: A meta-analysis", *Personnel Psychology*, vol 44, no. 1, pp. 1-26, Mar. 1991.

[50] G. Chen, D. Davis, C. Hauff, and G.J. Houben, "On the impact of personality in massive open online learning", *Proc. Conference on User Modeling Adaption and Personalization (UMAP '16)*, pp. 121-130, Jul. 2016, doi: 10.1145/2930238.2930240.

[51] D. Yang, R. Kraut, and C.P. Rosé, "Exploring the Effect of Student Confusion in Massive Open Online Courses", *Journal of Educational Data Mining*, vol. 8, no. 1, 2016.

[52] D. Yang, M. Wen, C.P. Rosé, "Weakly Supervised Role Identification in Teamwork Interactions", Proc. Annual Meeting of the Association for Computational Linguistics (ACL '15), pp. 1671-1680, Jul. 2015, doi: 10.3115/v1/p15-1161.

[53] Y. Shi, Z. Peng, and H. Wang, "Modeling Student Learning Styles in MOOCs", *Proc. ACM Conference on Information and Knowledge Management (CIKM '17)*, pp 979-988, Nov. 2017.

[54] S. Tang and Z.A. Pardos, "Personalized Behavior Recommendation: A case study of applicability to 13 courses on edX", *Adjunct Publication. Conference on User Modeling, Adaptation and Personalization (UMAP '17)*, pp. 165-170, July 2017.

[55] S.-S. Shen, H.-Y. Lee, S-W. Li, V. Zue, and L-S. Lee, "Structuring Lectures in Massive Open Online Courses (MOOCs) for Efficient Learning by Linking Similar Sections and Predicting Prerequisites", *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH '15)*, pp. 1363-1367, Sep. 2015.

[56] B.N. Green, C.D. Johnson, and A. Adams, "Writing narrative literature reviews for peer-reviewed journals: secrets of the trade", *Journal of chiropractic medicine*, vol. 5, issue 3, pp. 101-117, Autumn 2006.

[57] M.L. Pan, *Preparing literature reviews: Qualitative and quantitative approaches.*, Abingdom, UK: Routledge, 2016.

[58] M. Pérez-Sanagustín, M. Nussbaum, I. Hilliger, C. Alario-Hoyos, R.S. Heller, P. Twining, and C.C. Tsai, "Research on ICT in K-12 schools–A review of experimental and survey-based studies in computers & education 2011 to 2015", *Computers & Education*, vol. 104, pp. A1-A15, Jan. 2017.

[59] J.W. Gikandi, D. Morrow, and N.E. Davis, "Online formative assessment in higher education: A review of the literature", *Computers & education*, vol. 57, issue 4, pp. 2333-2351, Dec. 2011.

[60] D. Dermeval, R. Paiva, I.I. Bittencourt, J. Vassileva, and D. Borges, "Authoring Tools for Designing Intelligent Tutoring Systems: a Systematic Review of the Literature", *International Journal of Artificial Intelligence in Education*, pp. 1-49, Oct. 2017.

[61] A.N. Guz and J.J. Rushchitsky, "Scopus: A System for the Evaluation of Scientific Journals", *International Applied Mechanics,* vol. 45, no. 4, pp. 351-362, Oct. 2009.

[62] L. Steuten, G. van de Wetering, K. Groothuis-Oudshoom, and V. Retèl, "A systematic and critical review of the evolving methods and applications of value of information in academia and practice", *Phamacoeconomics*, vol. 31, no. 1, pp. 25-48, Dec. 2012

[63] M. Gonzalez-Loureiro, M. Dabic, and T. Kiessling, "Supply chain management as the key to a firm's strategy in the global marketplace: trends and research agenda", *International Journal of Physical Distribution & Logistics Management*, vol. 45, no. 1/2, pp. 159-181, 2015.

[64] C. Ye, J.S. Kinnebrew, G. Biswas, B.J. Evans, D.H. Fisher, G. Narasimham, and K.A. Brady, "Behavior prediction in MOOCs using higher granularity temporal information", *Proc. ACM Conference on Learning@ Scale (L@S '15)*, pp. 335-338, Mar. 2015, doi: 10.1145/2724660.2728687.

[65] J. Gardner and C. Brooks, "A Statistical Framework for Predictive Model Evaluation in MOOCs", *Proc. ACM Conference on Learning@ Scale (L@S '17)*, pp. 269-272, Apr. 2017.

[66] S. Crossley, L. Paquette, M. Dascalu, D.S. McNamara, and R.S. Baker, "Combining click-stream data with NLP tools to better understand MOOC completion", *Proc. International Conference on Learning Analytics & Knowledge (LAK '16)*, pp. 6-14, Apr. 2016, doi: 10.1145/2883851.2883931.

[67] Z.A. Pardos, and Y. Xu, "Improving efficacy attribution in a self-directed learning environment using prior knowledge individualization", *Proc. International Conference on Learning Analytics & Knowledge (LAK '16)*, pp. 435-439, Apr. 2016, doi: 10.1145/2883851.2883949

[68] J. Qiu, J. Tang, T.X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue, "Modeling and predicting learning behavior in MOOCs", *Proc. ACM International Conference on Web Search and Data Mining (WSDM '16)*, pp. 93-102, Feb. 2016, doi: 10.1145/2835776.2835842.

[69] W.H. Wu, Y.C.J Wu, C.Y. Chen, H.Y. Kao, C.H. Lin, and S.H. Huang, "Review of trends from mobile learning studies: A meta-analysis", *Computers & Education*, vol. 59, issue 2, pp. 817-827, Sep. 2012.

[70] D. Shah, "By the numbers: MOOCs in 2015", *Class Central*, https://www.class-central.com/report/moocs-2015-stats. 2015.

[71] S. Boyer, and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses", *Proc. International Conference on Artificial Intelligence in Education (AIED '15)*, pp. 54-63, Jun. 2015, doi: 10.1007/978-3-319-19773-9_6.

[72] D. Comer, "Learning how to teach... differently: Extracts from a MOOC instructor's journal", *Invasion of the MOOCs: The promise and perils of massive open online courses*, pp. 130-149, 2014.

[73] T.R. Liyanagunawardena, and S.A Williams, "Massive open online courses on health and medicine", *Journal of medical Internet research*, vol. 16, no.8, p. e191, Aug. 2014.

[74] M. Vitiello, S. Walk, D. Helic, V. Chang, and C. Guetl, "Predicting dropouts on the successive offering of a MOOC", *Proc. International Conference MOOC-MAKER (MOOC-MAKER '17)*, pp. 11-20, Nov. 2017.

[75] C.A. Coleman, D.T. Seaton, and I. Chuang, "Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification", *Proc. ACM Conference on Learning@ Scale (L@S '15)*, pp. 141-148, Mar. 2015, doi: 10.1145/2724660.2724662.

[76] B. Xu, and D. Yang, "Motivation classification and grade prediction for MOOCs learners", *Computational intelligence and neuroscience*, vol. 2016, no. 4, Jan. 2016.

[77] F. Jiang, and W. LI, "Who Will Be the Next to Drop Out? Anticipating Dropouts in MOOCs with Multi-View Features", *International Journal of Performability Engineering*, vol. 13, no. 2, pp. 201-210, Mar. 2017.

[78] C.G. Brinton, and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks", *Proc. IEEE Conference on Computer Communications (INFOCOM '15)*, pp. 2299-2307, May. 2015, doi: 10.1109/INFOCOM.2015.7218617.

[79] J. Arguello, and K. Shaffer, "Predicting Speech Acts in MOOC Forum Posts", *Proc. AAAI Conference on Web and Social Media (ICWSM '15)*, pp. 2-11, May 2015.

[80] D. Yang, D. Adamson, and C.P. Rosé, "Question recommendation with constraints for massive open online courses", *Proc. ACM Conference on Recommender systems (RecSys '14)*, pp. 49-56, Oct. 2014, doi: 10.1145/2645710.2645748.

[81] N.M. Hicks, D. Roy, S. Shah, K.A. Douglas, P. Bermel, H.A. Diefes-Dux, and K. Madhavan, "Integrating analytics and surveys to understand fully engaged learners in a highly-technical STEM MOOC", *Proc. Frontiers in Education Conference (FIE '16)*,

pp. 1-9, Oct. 2016, doi: 10.1109/FIE.2016.7757735.

[82] M.L. Bote-Lorenzo, and E. Gómez-Sánchez, "Predicting the decrease of engagement indicators in a MOOC", *Proc. International Conference on Learning Analytics & Knowledge (LAK '17)*, pp. 143-147, Mar. 2017, doi: 10.1145/3027385.3027387.

[83] D. Vu, P. Pattison, and G. Robins, "Relational event models for social learning in MOOCs", *Social Networks*, vol. 43, pp. 121-135, Oct. 2015.

[84] H. Drachsler and M. Kalz, "The MOOC and learning analytics innovation cycle (MOLAC): a reflective summary of ongoing research and its challenges", *Journal of Computer Assisted Learning*, vol. 32, issue 3, pp. 281-290, Mar. 2016.

[85] R. Al-Shabandar, A.J. Hussain, A. Laws, R. Keight, and J. Lunn, "Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses", *Proc. International Joint Conference on Neural Networks (IJCNN '17)*, pp. 713-720, May 2017.

[86] J.A. Greene, C.A. Oswald, and J. Pomerantz, "Predictors of retention and achievement in a massive open online course", *American Educational Research Journal*, vol. 52, issue 5, pp. 925-955, Oct. 2015.

[87] C.G. Brinton, S. Buccapatnam, M. Chiang, and H.V. Poor, "Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance", *IEEE Trans. on Signal Processing*, vol. 64, issue. 14, pp. 3677-3692, Jul. 2016, doi: 10.1109/TSP.2016.2546228.

[88] S. Boyer, and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses", *Proc. International Conference on Artificial Intelligence in Education (AIED '15)*, pp. 54-63, Jun. 2015, doi: 10.1007/978-3-319-19773-9_6.

[89] M. Klüsener, and A. Fortenbacher, "Predicting students' success based on forum activities in MOOCs", *Proc. International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS '15)*, vol. 2, pp. 925-928, Sep. 2015, doi: 10.1109/IDAACS.2015.7341439.

[90] J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction", *Proc. International Conference on Computer Science & Education (ICCSE '16)*, pp. 52-57, Aug. 2016, doi: 10.1109/ICCSE.2016.7581554.

[91] S.-H. Zhong, Y. Li, Y. Liu, and Z. Wang, "A computational investigation of learning behaviors in MOOCs", *Computer Applications in Engineering Education*, vol. 25, no. 5, pp. 693-705, May 2017.

[92] B. Hong, Z. Wei, and Y. Yang, "Discovering Learning Behavior Patterns to Predict Dropout in MOOCs", *Proc. International Conference on Computer Science & Education (ICCSE '17)*, pp. 700-704, Aug. 2017.

[93] 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)", *Proc. ACM Conference on Learning@ Scale*, pp. 245-248, Mar. 2015.

[94] B. Makhabel, *Learning Data Mining with R*, Birmigham, UK: Packt Publishing Ltd., 2015.

[95] A.F. Wise, Y. Cui, W. Jin, and J. Vytasek, "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling", *Internet and Higher Education*, vol. 32, pp. 11-28, Jan. 2017.

[96] V.N. Vapnik, *The Nature of Statistical Learning Theory*, New York, NY: Springer, 2010.

[97] M. Fei and D. Yeung, "Temporal models for predicting student dropout in massive open online courses", *Proc International Conference on Data Mining Workshop (ICDMW '15)*, pp. 256-263, Nov. 2015, doi: 10.1109/ICDMW.2015.174.

[98] J. Macina, I. Srba, J.J Williams, and M. Bielikova, "Educational Question Routing in Online Student Communities", *Proc. ACM Conference on Recommender Systems (RecSys '17)*, pp. 47-55, Aug. 2017.

[99] J.R. Quinlan, "Simplifying decision trees", *International Journal of Man-Machine Studies*, vol. 27, issue 3, pp. 221-234, Sep. 1987.

[100] J. Gardner, and C. Brooks, "Statistical Approaches to the Model Comparison Tasks in Learning Analytics", *Proc. Workshop on Methodology in Learning Analytics (MLA) and the Workshop on Building the Learning Analytics Curriculum (BLAC)*, Mar. 2017.

[101] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling", *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947-1958, Nov. 2003.

[102] R.N. Laveti, S. Kuppili, J. Ch, S.N. Pal, and N.S.C. Babu, "Implementation of Learning Analytics Framework for MOOCs using State-of-the-art In Memory Computing", *Proc. National Conference on E-Learning & E-Learning Technologies (ELELTECH '17)*, pp. 1-6, Aug. 2017.

[103] I. Rish, "An empirical study of the naive Bayes classifier", *Proc. IJCAI Workshop on Empirical Methods in Artificial Intelligence (IJCAI '01)*, vol. 3, no. 22, pp. 41-46, Aug. 2001.

[104] W. Lu, T. Wang, M. Jiao, X. Zhang, S. Wang, X. Du, and H. Chen, "Predicting Student Examinee Rate in Massive Open Online Courses", *Proc. International Conference on Database Systems for Advanced Applications (DASFAA '17)*, pp. 340-351, doi: 10.1007/978-3-319-55705-2_27.

[105] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001.

[106] 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", *Proc. European MOOCs Stakeholders Summit (EMOOCS '17)*, pp. 263-272, May 2017, doi: 10.1007/978-3-319-59044-8_31.

[107] C.E. Imrie, S. Durucan, and A. Korre, "River flow prediction using artificial neural networks: generalization beyond the calibration range", *Journal of Hydrology*, vol. 233, no. 1, pp. 138-153, Jun. 2000.

[108] A. Pérez-Lemonche, G. Martínez-Muñoz, and E. Pulido-Cañabate, "Analysing Event Transitions to Discover Student Roles and Predict Grades in MOOCs", *Proc. International Conference on Artificial Neural Networks*, pp. 224-232, Sep. 2017, doi: 10.1007/978-3-319-68612-7_26.

[109] Z. Wang, J. Zhu, X. Li, Z. Hu, and M. Zhang, "Structured Knowledge Tracing Models for Student Assessment on Coursera", *Proc. ACM Conference on Learning@ Scale (L@S '16)*, pp. 209-212, Apr. 2016, doi: 10.1145/2876034.2893416.

[110] F. Arabshahi, F. Huang, A. Anandkumar, C.T. Butts, and S.M. Fitzhugh, "Are You Going to the Party: Depends, Who Else is Coming?: [Learning Hidden Group Dynamics via Conditional Latent Tree Models]", *Proc. International Conference on Data Mining (ICDM '15)*, pp. 697-702, —Nov. 2015, doi: 10.1109/ICDM.2015.146.

[111] A.D. Mezaour, "Filtering Web Documents for a Thematic Warehouse Case Study: eDot a Food Risk Data Warehouse (ex-

tended)", *Proc. International IIS Intelligent Information Processing and Web Mining (IIPWM '05)*, pp. 269-278, Jun. 2005, doi: 10.1007/3-540-32392-9_28.

[112] C. Brooks, C. Thompson, and S. Teasley, "Towards A General Method for Building Predictive Models of Learner Success using Educational Time Series Data", *Proc. Workshops at the International Conference on Learning Analytics & Knowledge (LAK '14)*, Mar. 2014.

[113] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics", *Computer*, vol. 49, issue 4, pp. 61-69, Apr. 2016.

[114] J. Carletta, "Assessing Agreement on Classification Tasks: The Kappa Statistic", *Computational Linguistics,* vol. 22, no. 2, pp. 249-254, Jun. 1996.

[115] J.R. Landis and G.G. Koch, "The Measurement of Observer Agreement for Categorical Data", *Biometrics,* vol. 33, no. 1, pp. 159-174, Mar. 1977.

[116] D.S. Chaplot, E. Rhim, and J. Kim, "Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks", *Proc. International Conference on Artificial Intelligence in Education (AIED '15)*, pp. 7-12, Jun. 2015.

[117] R. Pelánek, "Metrics for evaluation of student models," *Journal of Educational Data Mining*, vol. 7, no. 2, pp. 1-19, 2015.

[118] R.F. Kizilcec, M. Pérez-Sanagustín, J.J. Maldonado, "Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses", *Computers and Education*, vol. 104, pp. 18-33, Jan. 2017.

[119] D.F. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: behavioural patterns", *Proc. International Conference on Education and New Learning Technologies (EDULEARN '14)*, pp. 5825-5834, Jul. 2014.

[120] T. Fawcett, "An introduction to ROC analysis", *Pattern recognition letters*, vol. 27, issue 8, pp. 861-874, Jun. 2006.

[121] D.R. Garrison, T. Anderson, and W. Archer, "Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education", *The Internet and Higher Education*, vol. 2, no. 2-3, pp. 87-105, 2000.

[122] G. Siemens, S. Dawson, and G. Lynch, "Improving the Quality and Productivity of the Higher Education Sector", *Policy and Strategy for Systems-Level Deployment of Learning Analytics*, Canberra, ACT: Society for Learning Analytics Research for the Australian Office for Learning and Teaching, 2013.

Pedro Manuel Moreno-Marcos is a PhD student and pre-doctoral researcher though a FPU fellowship in the Department of Telematics Engineering at the Universidad Carlos III de Madrid. He received his Bachelor in Telecommunications Technologies Engineering in 2015 as well as his Master Degrees in Telecommunication Engineering and Telematic Engineering, which were both obtained in 2017. All of them (bachelor and masters) were obtained at Universidad Carlos III de Madrid. His areas or research interest include learning analytics, Educational Data Mining and MOOCs (Massive Open Online Courses).

Carlos Alario-Hoyos is a Visiting Associate professor in the Department of Telematics Engineering at the Universidad Carlos III de Madrid. He received M.S. degree in Telecommunication Engineering and PhD in Information and Communication Technologies from Universidad of Valladolid, Spain, in 2007 and 2012 respectively. Carlos has received several awards for his work on educational technologies, including best paper award at EC-TEL 2013. He is author of more than 60 scientific publications and has participated in more than 20 research projects, in the regional, national and European levels. His skills and experience include research and development in MOOCs, social networks, collaborative learning, and evaluation of learning experiences, among others.

Pedro J. Muñoz-Merino is a lecturer and researcher at the Universidad Carlos III de Madrid, where he is the Director of the Master in Telematics Engineering. In 2009, he got his PhD in Telematics Engineering from the Universidad Carlos III de Madrid. He obtained his accreditation in May 2012 as Associate Professor by the ANECA agency from the Spanish Ministry of Education. Pedro has received several awards for his work on educational technologies. He is author of more than 90 scientific publications and has participated in more than 30 research projects, coordinating some of them with private companies. His skills and experience include research and development in learning analytics, educational data mining, evaluation of learning experiences, user studies, gamification or Intelligent Tutoring Systems.

Carlos Delgado Kloos received the PhD degree in Computer Science from the Technische Universitat Munchen and in Telecommunications Engineering from the Universidad Politecnica de Madrid. He is full professor of Telematics Engineering at the Universidad Carlos III de Madrid, where he is the director of the GAST research group, director of the UNESCO Chair on “Scalable Digital Education for All”, and vicepresident for Strategy and Digital Education. He is also the Coordinator of the eMadrid research network on Educational Technology in the Region of Madrid. He is the Spanish representative at IFIP TC3 on Education.