Informatics in Education, 2021, Vol. 20, No. 2, –

DOI: 10.15388/infedu.2021.09

Active Methodology, Educational Data Mining and Learning Analytics: A Systematic Mapping Study

Tiago Luís de ANDRADE, Sandro José RIGO,
Jorge Luis Victória BARBOSA

Graduate Program in Applied Computing, University of Vale do Rio dos Sinos
São Leopoldo/RS, Brazil
e-mail: tiago@unemat.br, rigo@unisinos.br, jbarbosa@unisinos.br

Received: September 2020

Abstract. Distance Learning has enabled educational practices based on digital platforms, generating massive amounts of data. Several initiatives use this data to identify dropout contexts, mainly providing teacher support about student behavior. Approaches such as Active Methodologies are known as having good potential to involve and motivate students. This article presents a systematic mapping aiming to identify current Educational Data Mining and Learning Analytics methods. Besides, we identify Active Methodologies' application to mitigate dropout in Distance Learning. We evaluated 668 papers published from January 2015 to March 2020. The results indicate a growing application of Educational Data Mining and Learning Analytics to identify and mitigate students' abandonment in Distance Learning. However, studies with Active Methodologies to minimize dropout and enhance student permanence are scarce. Some works suggest Active Methods as a possible complement of Learning Analytics in dropout.

Keywords: active methodology, educational data mining, learning analytics, dropout, distance education.

1. Introduction

Technology is helping to create a profound impact on education, transforming the way of knowledge transmission and the scope of teaching and learning. Distance Education is a teaching method that allows self-learning, in which teachers and students can be geographically distant and interact asynchronously (Heidrich *et al.,* 2018) through Virtual Learning Environments (VLE), where all the course content is presented, as well as interaction and evaluation resources are available (Ramos *et al.,* 2018).

As its acceptance and popularity have increased, a concern has plagued educational institutions and has become a reality: the high dropout rate. Periodically, surveys are conducted to collect information about the courses offered in distance education. Bra-

zilian Annual Census had an average dropout rate of 18.6% in 2010, 20.5% in 2011, 11.74% in 2012, and 16.94% in 2013 (Queiroga *et al.,* 2017). In 2015, in 40% of the institutions surveyed, the dropout rate was between 26% to 50% (Ramos *et al.,* 2017). In 2016, in 32% of the institutions surveyed, the average dropout rate was between 11% and 25% and, for another 13% of institutions, from 26% to 35% (Ramos *et al.,* 2018). In 2018, indications of up to 50% dropout were found in totally distance courses. This demonstrates that the teaching modality suffers from high dropout rates, and research for methods that help reduce these numbers is one of its main challenges (Queiroga *et al.,* 2019). When comparing with classroom teaching, students who have up to 25% of the course completed, the dropout rate in distance education is lower; from 25% to 50%, the rate is higher; more than 50%, the rates are similar. About the numbers, there is evidence of the superiority of dropout in distance education compared to classroom teaching (Oliveira and Bittencourt, 2020).

One of the prevention hypotheses verified in the distance education literature is the use of Active Methodologies for students prone to dropout identified through the techniques of Educational Data Mining (EDM) and Learning Analytics (LA). The Active Methodology induces collaborative activity, teamwork, critical sense development, and the ability to argue, making the student protagonist of his learning process using the most varied resources of virtual environments. According to Guo *et al.* (2018), an Active Methodology is an approach in which students participate in the learning process. They are encouraged to interact with colleagues to develop activities, collaborate for intellectual growth, and improve the performance of those involved.

Given the above, this systematic mapping aims to verify the use of Educational Data Mining (EDM) and Learning Analytics (LA) for the identification of students prone to dropout and to analyze the Active Methodologies integration to LA, in order to assist the teacher in teaching and collaborate in mitigating dropout. As specific objectives, we aim to: (i) identify EDM and LA techniques, algorithms, and applications aimed at VLE in the process of prediction, detection, diagnosis or monitoring of students; and (ii) identify aspects regarding the use of Active Methods in educational platforms integrated with the LA process to mitigate dropout and enhance the permanence in distance education.

With the focus on distance education, we hope to contribute so that this teaching modality remains part of a continuous improvement process, incorporating new technologies and methodologies, helping to mitigate the risks of dropout, and enhancing the course permanence.

This article is organized into five sections that includes the introduction. Section 2 presents the theoretical foundation with the relevant concepts of distance learning and the main teaching platforms, dropout in distance education, EDM and LA in distance education, and finally, Active Methodologies and distance education. Section 3 describes the research methodology and the questions that permeate this study. Section 4 reports the main results obtained for each research question. Finally, section 5 presents the final considerations of this work.

2. Literature Review

This section presents concepts of Distance Education, dropout, EDM, LA, and Active Methods. It is noteworthy that these themes are interdisciplinary and they are present in the speeches and practices of teachers in the teaching and learning process.

2.1. *Distance Educational and Virtual Learning Environment*

Distance education has assumed an important role in the students learning processes. It requires technological environments capable of managing a great number of activities involved in the learning process. Therefore, Distance education generates a large amount of data that can serve as raw material for research due to the high level of digital mediation (Maschio *et al.,* 2018; Cambruzzi *et al.,* 2015), such as access logs, various interactions with the system, messages in forums, among others (Silva *et al.,* 2015). However, a large part of this data has not been analyzed, which constitutes a significant gap for conducting research, given the amount of valuable information that can potentially be extracted from them (Rabelo *et al.*, 2017).

Romero and Ventura (2013) state that managing data is one of the biggest challenges facing educational institutions since it has grown exponentially. For Waheed *et al.* (2020), educational data has been substantiated as a multidisciplinary field of study involving several research disciplines, generating several terms associated with this educational data exploration, such as academic analysis, predictive analysis, learning analysis and, finally, educational data science. According to Ramos *et al.* (2018), the data extracted from virtual environments can indicate students' behavioral characteristics and, therefore, allow inferential and predictive analyzes based on technology.

In this perspective, online systems, such as Learning Management System (LMS), Content Management System (CMS), Massive Open Online Course Platforms (MOOC), Virtual Learning Environments (VLE), among other web-based educational systems, contribute to generate digital data that can be analyzed to assess student behavior and assist teachers to improve each other's performance (Waheed *et al.*, 2020; Isidro *et al.*, 2018). According to Kostopoulos *et al.* (2019b), previously inaccessible data about students can be easily extracted from learning management systems and transformed into useful knowledge.

When it comes to teaching platforms, VLE is an online educational platform that provides teachers with the insertion of content for student learning. According to Queiroga *et al.* (2017), the organization of the VLE allows the student and the teacher to systematically monitor what should be studied each week throughout a given discipline. In the same vein as VLE, MOOC is an online learning platform founded by Stanford University in August 2011 (Wang and Wang, 2019), capable of generating a large amount of data from the interaction of students in the learning process (Isidro *et al.*, 2018). This research will address the studies that used online educational platforms, not restricted to the two mentioned in this section.

2.2. *Dropout in Distance Education*

Distance education is an alternative for accessing personal, professional, and academic learning programs. However, it can have associated problems, such as high dropout rates. Dropout is considered one of the most severe problems affecting educational institutions and a matter of concern for academic managers. School dropout is classified into three groups:

(i) Economic – impossibility to stay in the course due to socio-economic issues.
(ii) Vocational – the student does not identify with the course.
(iii) Institutional – abandonment due to failure in the initial disciplines, previous deficiencies in previous contents, inadequacy in the study methods, difficulties in relationships with colleagues or with members of the institution (Manhães *et al.,* 2011).

Identifying the reason that leads the student to dropout is important for the educational institution and the teacher so that it is possible to provide the necessary conditions that reduce or eliminate it. Therefore, analyzing the students' performance is an important factor because it can represent the degree of difficulty in learning the content and thus determine risk of failure and possible dropout. Once identified, it is possible to propose proactive actions among the actors, such as supporting underperforming students with seminars, intervention programs, workshops, and additional learning material, resulting in lower dropout rates (Manhães *et al.,* 2011; Kostopoulos *et al.,* 2018a; Kostopoulos *et al.,* 2018b). For that, data analysis methods and tools are needed to observe students' behavior to assist stakeholders in decision making (Silva *et al.,* 2015).

In this sense, considerable emphasis was placed on predicting performance, an essential and complex task, which has been studied in detail to detect the success or failure of distance learning students. The difficulty lies in the fact that this prediction is influenced by factors such as academic background, social environment, student demographic characteristics, and the educational environment in question (Kostopoulos *et al.,* 2019a; Tomasevic *et al.,* 2020).

Thus, the concern with learning environments is once again perceived, since student involvement and interaction with the teaching platform is an important factor in performance forecasting and learning analysis. Tools like Moodle* can track the amount of access to learning materials (Tomasevic *et al.,* 2020). The same can be said for MOOC courses, which have a high dropout rate (Wang *et al.,* 2017; Isidro *et al.,* 2018), a potential factor that hinders its development. Therefore, predicting whether a student will dropout a course is of great value for teaching platforms.

In this scenario, EDM and LA are presented as an alternative for the treatment and discovery of knowledge in the bases generated by the students' information on educational platforms. In this way, it has been establishing itself as a strong and consolidated line of research, which has excellent potential for improving the quality of distance learning (Baker *et al.,* 2011; Queiroga *et al.,* 2019).

* https://moodle.org/

2.3. *EDM and LA in Distance Education*

While the large volume of data enabled a more accurate study on dropout, identifying the relevant information in the database is not an easy task, as it requires technical knowledge about the databases on the teaching platforms and appropriate analytical tools (Silva *et al.,* 2015). The implementation of data mining and machine learning methods in the educational field has boosted the use of data gathered in different teaching contexts, leading to the development of two interrelated areas: Educational Data Mining (EDM) and Learning Analytics (LA) (Kostopoulos *et al.,* 2019a).

EDM is a growing area of scientific research and is closely linked to LA. According to Baker and Yacef (2009) and Kostopoulos *et al.* (2019a), EDM is characterized as an efficient interdisciplinary research area to unravel knowledge of educational data, forming an integral element of the learning process of students, educators, and educational institutions. LA is a fast-growing research field, mainly focused on the development and application of processes and tools to collect, explore and analyze large amounts of data, to understand students' learning behavior better, help teachers to give better support and appropriate interventions and, finally, improve the quality of learning and teaching, as well as educational outcomes.

Looking for additional definitions, Queiroga *et al.* (2017) state that data mining emerges as an alternative for the treatment and discovery of knowledge within this large volume of data generated by VLEs. For Baker *et al.* (2011), EDM is defined as the research area whose primary focus is the development of methods to explore sets of data collected in educational environments. Similarly, Romero and Ventura (2007) define EDM as the application of mining techniques to data from online education platforms or environments. Santos *et al.* (2016) add that it is possible to understand students' learning and the context involved in this process more effectively and appropriately through EDM. In the same vein, Romero *et al.* (2008) and Silva *et al.* (2015) state that it is an interesting inductive approach that creates models to automatically discover individual information present in student data that can improve learning. As information is obtained from the foundations of the teaching platforms, learning analysis is becoming crucial in the student assessment process, making it more efficient and accurate (Tomasevic *et al.,* 2020).

About LA, Brito *et al.* (2019) state that it has attracted the attention of the scientific community that works with educational technologies, as it provides a more effective way for teachers to track student performance and involvement in educational platforms. However, for the application, accurate data must be provided for processing and analysis since erroneous data can easily lead to misinterpretation and inaccuracy of the obtained results (Tomasevic *et al.,* 2020).

As highlighted, EDM and LA share the same objective: to improve the teaching and learning process by improving the evaluation processes, understanding the problems of education and planning interventions (Siemens and Baker, 2012). In this sense, despite this understanding, the results of this mapping come from EDM and LA's junction since the studies show few differences found and that both use similar techniques and methods, such as classification, grouping, regression, and visualization.

2.4. *Active Methodology and Distance Education*

Education is currently undergoing a transformation process, as new teaching and learning techniques and methods are being introduced. This has been a challenge for teachers and students, as it imposes the flexibility of opening to the new and the ability to learn (Chandrasekaran *et al.,* 2016).

The Active Methodology is considered important in the learning process since it involves students to participate in the construction of knowledge actively and changes the role of the teacher, who was previously a transmitter of content and information for a learning facilitator (Chandrasekaran *et al.,* 2016). It is important to highlight that the use of Active Methods in teaching induces aspects of active learning, which includes other concepts, such as collaborative learning and cooperative learning. In active learning students learn the content from the development of activities defined by the teacher, responsible for supervising and proposing discussions and challenges, and performed through collaborative or cooperative learning, which involves two or more participants.

This methodology is in the process of adoption and expansion in classroom and distance education since it is possible to find several research types on the subject. In a historical context, according to Chandrasekaran *et al.* (2016), at the beginning of the 19th century, a computer-mediated collaborative learning system was developed to support distance education, which was characterized by the online provision of shared workspaces and interactive chat. Over the years and the continually improving, of the most remarkable recent advances, in 2007, the Wiki was launched, a computer system that promoted collaboration between students of distance education through online discussions and sharing of experiences. In 2017 Lima and Siebra (2017) developed a tool called CollabEduc that aimed to motivate collaboration between participants throughout educational activities in distance education, through the concern with high dropout rates.

In this sense, the Active Methodology every day shows to be efficient in the teaching and learning process. According to Guo *et al.* (2018), it enhances intellectual growth and improves the performance of students involved in activities, as well as stimulating the relationship between colleagues. In this perspective, Gokhale (1995) states that Active Methodology gives the student the responsibility to learn and collaborate in the other's learning. For Tjhin *et al.* (2017), it stimulate students to have good management of study time and the ability to self-learn.

According to Chandrasekaran *et al.* (2016), students feel comfortable studying through the Active Methodology, as this methodology offers the opportunity to express individual experiences and share ideas in groups, promotes the development of social skills for those who have difficulties in learning practices centered only on the teacher, assigns much of the responsibility for learning, and enriches students experience with aspects of critical thinking and problem-solving. Still, according to Chandrasekaran *et al.* (2016), the research showed that students learn in different ways when working as a team, and one of them is precisely learning from the experiences of other colleagues.

There are several examples of Active Methodologies. Problem Based Learning (PBL) occurs from the study of a subject and group discussion in search of a solution

to the problems encountered. Hybrid Teaching mixes learning via the educational and classroom platform in the classroom. Peer Instruction Learning involves the student in studying theories at home and in discussing topics defined by the teacher in the classroom. Gamification turns classroom situations into problem-solving creatively through game elements in the teaching process. In Just-in-Time-Teaching, the teacher makes available before class the material for the student to study and answer a questionnaire, which will serve to identify the most difficult topics and prepare the content to be taught in the classroom. In this context, approaches as Just-for-you are also mentioned as options to increase intelligent personalization and pedagogical support for the students (Thomas *et al.,* 2016).

In distance education, according to Leite and Ramos (2017), the Active Methodology provides the student with interaction with the teacher, the colleague, the content and technology in the VLEs, a factor that can enhance the teaching and learning process and interaction if learners know-how to virtually explore the environment through the exercise of curiosity, critically and reflectively. In this teaching modality, in practical terms, while the student facilitates learning through access to relevant materials, the teacher can select the material according to the needs and achieve the desired results.

Also according to Leite and Ramos (2017), the use of Active Methodologies in distance education, considered as pedagogical innovation, goes beyond the traditional methodology in which the teacher is the transmitter of knowledge, as it offers students the possibility of seeking solutions for different situations, and with that, develop skills such as autonomy, interaction, cooperation, collaboration and commitment to learning itself. Thus, the VLEs developed for distance education courses must be attentive to the adoption of an active learning methodology so that there is more possibility for the student to build knowledge in a meaningful way.

Lima and Siebra (2017) used the methodology in question to mitigate the possibility of dropout and enhance the permanence in courses offered in distance education. Also, the referred authors cited or made use of this methodology in the context of assisting the teaching and learning process of students and teachers. Therefore, we consider that Active Methodology can be applied to VLEs and this teaching method can be an important ally in combating students' dropout in distance education, since, given the characteristics reported, it encourages interaction between the actors and the educational platform, and collaborates with collaborative and pedagogical practices.

3. Methodology for Mapping Process

The methodology used for this study was the Systematic Mapping of Literature proposed by Petersen *et al.* (2015), whose execution consisted of the following steps: the formulation of research questions, definition of the criteria for inclusion and exclusion of studies, search and selection of articles to be analyzed, evaluation of studies, and, finally, data collection. Afterward, the analysis and presentation stage was performed in the form of graphs, tables, and descriptions, supporting the interpretation of results and discussions.

Table 1
Research questions of the study

ID	Question
RQ1	Has dropout in distance education been an object of study?
RQ2	What techniques or methods of EDM and LA are used to predict, detect, diagnose, or monitor dropout in distance education?
RQ3	What were the computational tools used to address dropout in distance education?
RQ4	What were the algorithms used to deal with dropout in distance education?
RQ5	What were the attributes used in dropout studies in distance education?
RQ6	What is the level of education of the target audience in studies on dropout in distance education?
RQ7	Was an Active Methodology used to mitigate dropout in distance education?

3.1. *Research Questions*

For the research process, we defined the research questions presented in Table 1.

3.2. *Search Strategy*

As a strategy to find the relevant studies, an automatic search in electronic bases and a manual search in conferences were performed to ensure that the largest number of studies could be verified.

The following electronic databases were searched:

- Institute of Electrical and Electronics Engineers (IEEE).
- Science Direct by Elsevier.
- Association for Computing Machinery (ACM).
- Google Scholar search tool.

A search was carried out at the Brazilian Symposium on Computers in Education (SBIE), which has an H5 index of 15. This research source was included given its significant contribution in EDM and LA in Brazil (Rodrigues *et al.,* 2018) and also to identify possible social and cultural factors impacting this specific research area.

The search in the electronic databases was performed using expressions from keywords, including synonyms or related words to compose the terms. In English, the string used was (‘active learning’ OR ‘Active Methodology’) AND (‘edm’ OR ‘education data mining’ OR ‘educational data mining’) AND (‘dropout’ OR ‘evasion’) AND (‘distance learning’ OR ‘distance teaching’ OR ‘distance education’).

3.3. *Result Filters*

The articles were selected according to the following Inclusion Criteria: studies focusing on dropout in distance education; studies that used EDM and LA techniques; studies

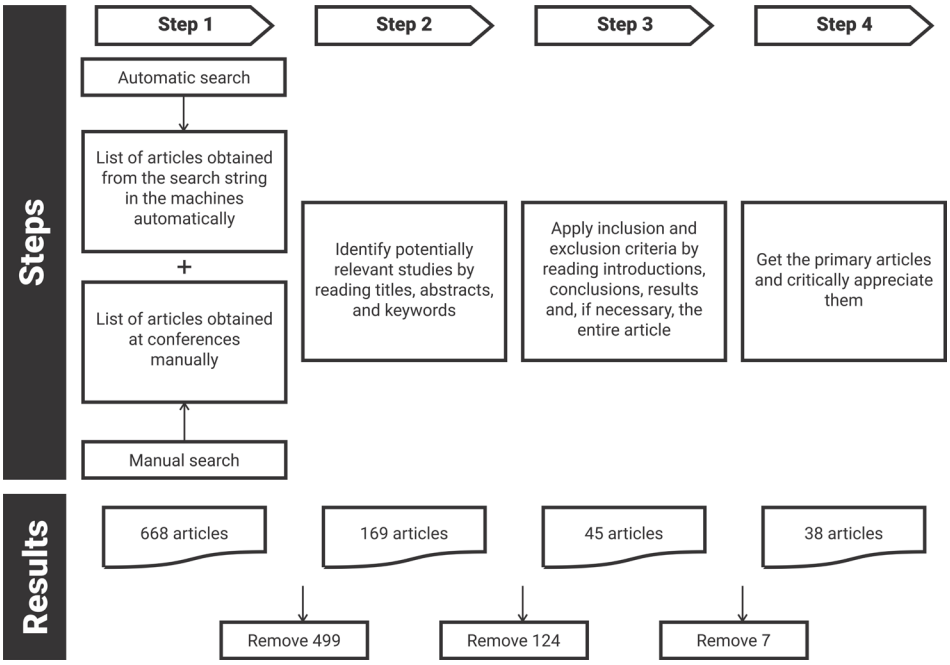

Fig 1. Results by stages of the study selection process.

that addressed the use of online educational platforms; studies that addressed training courses at the secondary, undergraduate or graduate levels; studies that addressed the use of tools to apply EDM and LA techniques; studies written in English and Portuguese; finally, studies published between January 2015 and March 2020.

In contrast, the Exclusion Criteria were: studies whose focus was dropout in classroom education; studies that did not deal with EDM or LA technical dropout in distance education; studies that addressed the hybrid method, only 100% distance courses are acceptable; dissertations, theses, and books were excluded.

Fig. 1 represents the steps taken to select the articles according to the search. The first step consisted of searching articles in the databases and conferences automatically and manually and eliminating duplicates, resulting in a total of 668 articles returned. Stage 2 consisted of identifying potentially relevant studies based on the analysis of the title, abstract, and keywords. In this stage 499 articles were discarded, and 169 selected for the next stage. In stage 3, the selected studies were reviewed by reading the introduction, results, and conclusions, applying the inclusion and exclusion criteria. If the reading of the previous items was not sufficient, the study was read in full. In this stage, 124 articles were discarded, and 45 selected for the next stage. Finally, in step 4, a list of 38 articles was obtained to be critically assessed, extracting the relevant data from each work.

Table 2 shows the number of articles initially obtained and those selected for reading by database or conference.

Table 2
Quantitative of articles obtained and selected in the study

Database/Conference	Return articles	Percentage of return articles	Selected articles	Percentage of selected articles
IEEE[1]	363	54.34%	12	31.58%
Elsevier Science Direct[2]	58	8.68%	4	10.53%
ACM[3]	56	8.38%	10	26.31%
Google Scholar[4]	166	24.85%	3	7.89%
SBIE[5]	25	3.74%	9	23.68%
TOTAL	668	100%	38	100%

[1] https://www.ieee.org/index.html
[2] https://www.elsevier.com/
[3] http://dl.acm.org/
[4] https://scholar.google.com.br/
[5] https://cbie.ceie-br.org/

4. Results and Discussion

This section summarizes the results obtained from the 38 primary articles analyzed in our research, considering each of the research questions.

RQ1: *Has dropout in distance education been an object of study?*

Dropout has been the subject of worldwide study, both in classroom and distance learning. The techniques of Data Mining and Learning Analysis have helped in the prediction, detection, diagnosis, or monitoring of students. Table 3 presents the database or conference and authors of the 38 articles that support the analysis. The objectives, results, and conclusions of the works with the most relevant contributions are shown after Table 3 to answer the question.

Kostopoulos *et al.* (2018b) investigate the efficiency of semi-supervised learning algorithms for data mining and machine learning capable of predicting student dropout rates in a distance course before it happens. The results indicated that the Naive Bayes algorithm had the best performance, with an accuracy of 66.26% in pre-university data (demographic and previous experiences) and 84.56% in academic data (first two written evaluations and presence), favoring the development of personalized learning strategies, increasing retention rates and improving the quality of education.

Brito *et al.* (2019) present the Dropout Risk Report tool developed to detect students' dropout risks through Moodle data, such as cognitive, social, and behavioral. According to the authors, the results made it easier for teachers, tutors, and educational directors to obtain more accurate information and indicators of access, performance, and interactions, to rescue students who had never accessed or abandoned the teaching platform, and to intervene immediately in cases identified with the risk of dropout.

Table 3
Articles selected for the research that address dropout in distance education

Database	Authors
IEEE	Kostopoulos *et al.* (2018b); Brito *et al.* (2019); Brandão *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); Liang *et al.* (2016); Mishra and Mishra (2018); Cobos and Olmos (2019); De La Peña *et al.* (2018); Macedo *et al.* (2019); Isidro *et al.* (2018); Wang and Wang (2019)
Elsevier	Oeda and Hashimoto (2017); Waheed *et al.* (2020); Heidrich *et al.* (2018); Tomasevic *et al.* (2020)
ACM	Islam *et al.* (2019); Chen and Zhang (2017); Whitehill *et al.* (2017); Wang *et al.* (2017); Kang and Wang (2018); Imran *et al.* (2019); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Borrella *et al.* (2019)
Google Scholar	Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a)
SBIE	Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Dos Santos and Falcão (2017); Rabelo *et al.* (2017); Queiroga *et al.* (2017); Ramos *et al.* (2018); Queiroga *et al.* (2019)

Brandão *et al.* (2019) apply EDM techniques to Moodle data to verify the performance of participants in distance learning courses and the dropout rate. According to the results obtained, Adaboost presents the best performance and K-means the worst performance. As for the fee, 25% of the participants dropped out for the following reasons: lack of time, technical difficulties with the platform, and the understanding that the chosen course is complicated.

Ortigosa *et al.* (2019) present the SPA tool (Sistema de Prediccioen de Abandono, dropout prediction system in Spanish), which issues a dropout risk alert to the predictive models developed based on the C5.0 algorithm. The results showed a calculation of more than 117 thousand risk points of 5700 students, which allowed for intervention and 13 thousand retention actions.

De Almeida Neto and Castro (2015) investigate dropout situations through Association Rules in data from students of courses offered on the ColabWeb platform. The results indicated that the APriori-Inverse Algorithm identified 205 dropout situations and that, with the application, only seven dropouts have materialized.

Liang *et al.* (2016) present a dropout prediction model based on behavioral learning data from students on the XuetangX platform. The results showed an 88% accuracy rate, helping the teacher to identify students with a high probability of dropping out.

Mishra and Mishra (2018) identify students predisposed to evade courses through the Random Forest Algorithms, CART and C4.5, applied to the Xeutang.com platform data logs and proposed the development of the PCA – Principal Component Analysis algorithm to track the student progress and act as an alert for the teacher to encourage the improvement of student performance in the course.

Cobos and Olmos (2019) propose a tool for generating predictive models of course completion and dropout, with several machine learning algorithms in the MOOC's data. In 7 platforms used, 18 thousand models were generated. Highlight for the Bayesian

Generalized Linear Model algorithm, accurately predicting the conclusion more significant than the dropout measured by the Stochastic Gradient Boosting algorithm.

De La Peña *et al.* (2018) propose the use of data mining techniques, specifically Logistic Regression, to predict whether the student would abandon the course based on the grades of activities obtained by each stored in Moodle. The results obtained from applying the Logit Act tool to more than 100 students in 5 courses were slightly better than the existing proposals in terms of accuracy, especially in the crucial weeks of the semester.

Isidro *et al.* (2018) propose a method to predict the risk of student dropout in a MOOC through machine learning techniques in attempting and solving exercises and the time used to watch video lessons. Algorithms like Naive Bayes, Adaboost, Support Vector Machine (SVM), LSTM, and CART were used, in which the best results were obtained by the last mentioned.

Wang and Wang (2019) present the E-LSTM algorithm to interpret student behavior based on the time interval between activities and interactions and demonstrate whether they are likely to evade the course offered at MOOC. According to the authors, when compared to Logistic Regression, SVM, Decision Tree, Gradient Boosting and Random Forest, E-LSTM had the best performance, which means that it has a strong ability to predict dropout.

Waheed *et al.* (2020) present a Deep Artificial Neural Network (Deep ANN) applied to student demographics, activities and click flow in VLE, to predict the risk of dropout and provide measures for early intervention in these cases. The results show that Deep ANN achieved an accuracy rate of around 84% to 93%, better than the Logistic Regression and SVM techniques, useful for identifying students with learning difficulties and providing additional support in activities.

Tomasevic *et al.* (2020) perform the comparison of supervised machine learning techniques based on similarity, model and probabilistic to discover students prone to dropout and predict their future achievements, such as the final exam grade. The results obtained demonstrated that the Artificial Neural Networks algorithm was more accurate, according to student interaction and performance data.

Chen and Zhang (2017) propose an unsupervised learning system for detecting dropout based on student behavioral data in the MOOC. Inspired by statistical studies that relate behavior and dropout, it achieved high effectiveness in finding students who abandon the course and suggest alternatives for prevention, such as more chances to do activities, extend the delivery period and encourage participation in discussion forums.

Wang *et al.* (2017) propose a model called ConRec Network capable of predicting whether students will abandon courses, automatically extracting characteristics from the raw data of the MOOC. The results demonstrated to be efficient in the extraction of the data since it eliminates inconsistency, saves time and human effort, capable of solving problems of prediction of the withdrawal in MOOC.

Kang and Wang (2018) examine dropout rates, identifies patterns of students at high risk of dropout and proposes predictive models through data mining. The results obtained with the Logistic Regression technique demonstrated an accuracy rate of 81.8% for unbalanced data and 75.9% for balanced data, reducing enrollment dropout by 3%.

Imran *et al.* (2019) investigate the performance of various architectures of Deep Artificial Neural Network to develop the model of prediction of student dropout using machine learning, varying the number of layers and neurons. The results obtained in the dropout forecast were: for 3 layers and 256 neurons, the rate was 99.52%; for 5 layers and 1024 neurons, it was 97.46%; for 7 layers and 64 neurons, the rate was 99.80%.

Kostopoulos *et al.* (2015) study whether semi-supervised techniques Self-Training, Co-Training, Democratic Co-Training, Tri-Training, RASCO and Rel-RASCO, classified with Naive Bayes and C4.5, can be useful in predicting school dropout in the distance higher education. The results of the experiments presented the Tri-Training method with C4.5 as the best accuracy rate, varying between 53.26% and 75.29%.

Wu *et al.* (2019) propose a Deep Artificial Neural Network model called CLMS-Net, which combines Convolutional Neural Network, Long Short-Term Memory Network and Support Vector Machine for the automatic extraction of data related to student behavior. The result demonstrated the efficiency of the proposed model in the automatic extraction of characteristics, generating savings in time and labor, avoiding inconsistency in manual extraction, and helping to predict students' dropout.

Borrella *et al.* (2019) propose the development of a predictive model for MOOC with machine learning to identify students at high risk of dropout, according to click flow data on the teaching platform, and subsidize the intervention by email to motivate students. The model was able to predict four of the five actual dropouts in the courses. However, the intervention did not affect reducing the rate.

Kostopoulos *et al.* (2018a) present a set of classification and regression algorithms to predict student performance in final exams in a distance course and proposes an algorithm combining REPTree and M5 'Rules. The results demonstrated the accuracy of 82.25% in the identification, in which it excelled in both methods, so the educators could apply intervention strategies before the student evaded the course.

Santos *et al.* (2015) identify using a predictive model the discouraged student in a VLE using data mining, precisely the Decision Tree technique, and Scherer's definitions for the specification of discouragement. The results pointed out the success rate of 91% of students prone to discouragement, and consequently, the dropout rate.

Silva *et al.* (2015) present a predictive model for diagnosing dropout in VLE based on student interactions in discussion forums, to serve as a starting point for stakeholders (teachers, tutors, managers) in decision making. According to the authors, the results of the Decision Tree technique had the best performances, with an accuracy rate above 73%, significant for the data sets used. The highlight for the J48 algorithm obtained the best performance, despite a small difference for the Naive Bayes algorithm.

Ramos *et al.* (2017) present a predictive model called CRISP-EDM capable of predicting students' dropout based on the variables that make up the Transactional Distance (TD). The results showed that the Logistic Regression method presented the best results, with a success rate higher than 89%, helping the teacher or tutor in preventive action and reversing possible students with dropout trends.

Rabelo *et al.* (2017) apply Data Mining techniques through decision trees on Moodle data to predict student success or failure during the course, based on participation, inter-

action, and performance on the platform. The results obtained demonstrated an accuracy between 93.9% and 96.5% of precision. However, it is not necessary to wait for the end of the discipline to know the final performance, contributing to proposals for actions of conduct adjustment during the teaching and learning process collaborating to decrease dropout rates.

Queiroga *et al.* (2017) present an approach for detecting students at risk of dropping out by counting interaction and attributes derived from VLE. For that, predictive models based on algorithms were tested and evaluated in 2 different scenarios:

(i) Training and evaluation within the same course.
(ii) Training with data from 3 courses and evaluation with the remaining data.

The results obtained were satisfactory, in that the models generated using machine learning algorithms presented better performances than the model based on means and standard deviations of the weekly interactions; and the insertion of derived variables with greater granularity (count of daily interactions) helped to improve the performance of the models compared to previous experiments.

Ramos *et al.* (2018) present the comparative analysis of five classifiers used in machine learning, evaluating its use in defining the predictive model of student dropout from a set of interaction data and the structure of the virtual environment. According to the analysis of the classifier based on Logistic Regression, of the 9,777 cases classified as not evaded by the algorithm, the model was right for 9,035 (92.4%). For those classified as dropouts, the model hit 1,196 out of 1,663 cases (71.92%).

Queiroga *et al.* (2019) present an approach for detecting students at risk of dropout using a genetic algorithm created to optimize classifiers, aiming to assist in the prediction task, based on the count of student interactions within the VLE and derived attributes. In terms of accuracy, in the first 25 weeks of the course, the results were precisely 6.6% higher than those obtained by traditional methods. In the 50-week prediction, the difference was 4.7%, with the genetic algorithm again with the best result. In the 75-week prediction, the difference was 1.9%. Therefore, the Genetic Algorithm obtained higher hit rates in practically all scenarios.

After the presentation of the works, there is a concern with alarming dropout rates in distance learning. Although identifying the causes is extremely difficult, some factors are cited and identified through the techniques of EDM and Learning Analytics as those that most contribute to dropouts, such as social interaction, assessment methods, and the expectation of frustrated learning. Also important are the factors that indicate the persistence in this type of teaching, according to Borrella *et al.* (2019): social belonging, motivation, satisfaction, and self-regulation. Therefore, according to Kostopoulos *et al.* (2019a), EDM and LA are important allies in the prevention and detection of students at risk, supporting specific intervention strategies.

In order to qualify the analysis, Fig. 2 shows the number of primary articles selected in English and Portuguese, from international databases and the cited conference, by year of publication.

Table 4 presents the list of countries, the number, and references of the articles analyzed on the topic, identified by the authors' home institution, the place of application,

or direct citation to the country in the text. About Brazil, 5 articles were published in international databases and 9 in the Brazilian Symposium on Computers in Education (SBIE).

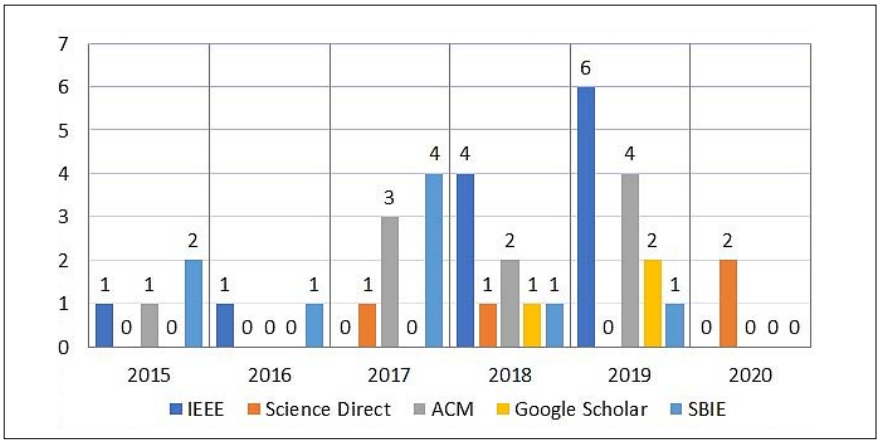

Fig 2. Articles selected by year of publication.

Table 4
List of articles by country

Country	Number of articles	References
Brazil	14	Brito *et al.* (2019); Brandão *et al.* (2019); De Almeida Neto and Castro (2015); Macedo *et al.* (2019); Heidrich *et al.* (2018); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Dos Santos and Falcão (2017); Rabelo *et al.* (2017); Queiroga *et al.* (2017); Ramos *et al.* (2018); Queiroga *et al.* (2019)
China	6	Liang *et al.* (2016); Wang and Wang (2019); Oeda and Hashimoto (2017); Chen and Zhang (2017); Niu *et al.* (2018); Wu *et al.* (2019)
Greece	5	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a)
Spain	4	Ortigosa *et al.* (2019); Cobos and Olmos (2019); De La Peña *et al.* (2018); Isidro *et al.* (2018)
USA	4	Whitehill *et al.* (2017); Kang and Wang (2018); Imran *et al.* (2019); Borrella *et al.* (2019)
Saudi Arabia	2	Waheed *et al.* (2020); Islam *et al.* (2019)
Índia	1	Mishra and Mishra (2018)
Serbia	1	Tomasevic *et al.* (2020)
Singapore	1	Wang *et al.* (2017)

RQ2. *What techniques or methods of EDM and LA are used to predict, detect, diagnose, or monitor dropout in distance education?*

First, it is important to define the concept of EDM and LA techniques and methods. Similarly treated, we consider techniques or methods to be the specification of the standards that we want to find through analysis, and that interest us in research, such as, for example, the number of students who repeat a given course subject. This data is only possible through the use of a technique or method of analysis, in this case, statistics.

In recent studies, several EDM techniques deal with dropout in distance education, from prediction to monitoring student performance, in order to assist teachers and educational managers in decision making. According to Kostopoulos *et al.* (2018a), the prediction has become an essential and challenging topic in the educational field, considered one of the most interesting and studied aspects of EDM. In the context of learning analysis, alert systems can be built from identification through the use of techniques of EDM (Ortigosa *et al.*, 2019).

Based on the primary articles studied, Table 5 presents the techniques or methods, the quantity, and the list of articles they used. The highlight for the classification and prediction, used in 32 and 25 articles, respectively, in the study of dropout in distance education for the learning behavior and performance of students on the teaching plat-

Table 5
List of articles by techniques or methods used

Techniques/Methods	Number of articles	References
Classification	32	Kostopoulos *et al.* (2018b); Brandão *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); Liang *et al.* (2016); Cobos and Olmos (2019); Isidro *et al.* (2018); Wang and Wang (2019); Waheed *et al.* (2020); Heidrich *et al.* (2018); Tomasevic *et al.* (2020); Islam *et al.* (2019); Chen and Zhang (2017); Whitehill *et al.* (2017); Wang *et al.* (2017); Kang and Wang (2018); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Borrella *et al.* (2019); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Dos Santos and Falcão (2017); Rabelo *et al.* (2017); Queiroga *et al.* (2017); Ramos *et al.* (2018); Queiroga *et al.* (2019)
Prediction	25	Kostopoulos *et al.* (2018b); Mishra and Mishra (2018); Cobos and Olmos (2019); De La Peña *et al.* (2018); Isidro *et al.* (2018); Wang and Wang (2019); Waheed *et al.* (2020); Heidrich *et al.* (2018); Tomasevic *et al.* (2020); Whitehill *et al.* (2017); Wang *et al.* (2017); Kang and Wang (2018); Imran *et al.* (2019); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Borrella *et al.* (2019); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2019a); Silva *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Queiroga *et al.* (2017); Ramos *et al.* (2018); Queiroga *et al.* (2019)
Clustering	6	Brandão *et al.* (2019); Cobos and Olmos (2019); Macedo *et al.* (2019); Oeda and Hashimoto (2017); Islam *et al.* (2019); Dos Santos and Falcão (2017)
Regression	2	Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a)
Summary/ Visualization	1	Islam *et al.* (2019)
Not Specified	1	Brito *et al.* (2019)

form, enables the achievement of positive results, timely and effective interventions. It is worth mentioning that in several articles, one or more techniques or methods are used in parallel. Kostopoulos *et al.* (2019b) use Classification, Prediction and Regression techniques to predict student performance through the development of a semi-supervised learning algorithm; Cobos and Olmos (2019) use Classification, Prediction, and Clustering in the tool developed to predict students who will or will not complete the course.

Corroborating with Waheed *et al.* (2020), several studies implement machine learning techniques to analyze behavior and predict students at risk of dropping out. However, there is no consensus among researchers as to which combination of techniques could produce the best results since the superiority of one model over another in predicting dropout cannot be affirmed in general for several reasons, such as problem specification and the type and characteristics of data to be analyzed (Imran *et al.,* 2019).

RQ3. *What were computational tools used to address dropout in distance education?*

As previously mentioned, for the functioning of distance education, computational tools are necessary to provide materials, to mediate communication between instructors and

Table 6

Tools used in articles for data collection and storage

Tools	Number of articles	References
Moodle	11	Brandão *et al.* (2019); Ortigosa *et al.* (2019); De La Peña *et al.* (2018); Macedo *et al.* (2019); Heidrich *et al.* (2018); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Rabelo *et al.* (2017); Queiroga *et al.* (2019)
Hellenic Open University (HOU)	4	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a)
EdX MOOC	3	Cobos and Olmos (2019); Isidro *et al.* (2018); Chen and Zhang (2017)
XuetangX MOOC	2	Liang *et al.* (2016); Wang *et al.* (2017)
MitX MOOC	2	Imran *et al.* (2019); Borrella *et al.* (2019)
HarvardX MOOC	2	Whitehill *et al.* (2017); Imran *et al.* (2019)
MOOC	2	Wang and Wang (2019); Oeda and Hashimoto (2017)
Open University Learning Analytics (OULA)	2	Waheed *et al.* (2020); Tomasevic *et al.* (2020)
Index of Learning Style Questionnaire (ILSQ)	1	Heidrich *et al.* (2018)
Universitas-XXI	1	Ortigosa *et al.* (2019)
ColabWeb	1	De Almeida Neto and Castro (2015)
ICourse163	1	Niu *et al.* (2018)
KDDCup	1	Wu *et al.* (2019)
Oracle	1	Mishra and Mishra (2018)
Kell	1	Kostopoulos *et al.* (2015)

students, and to promote the teaching and learning process, capable of generating and storing a huge amount of significant data.

To answer this question, we categorized the responses in three dimensions according to subjects found in the articles studied:

(i) Tools used to collect and store data.
(ii) Tools to extract and analyze stored data.
(iii) New tools developed by the authors of the primary articles to study dropout.

Table 6 presents data from the first dimension, in which 35 of the 38 primary articles disclosed the tools used as a data source, ordered by the largest number of uses. Of the 15 tools reported in 11 articles, the most cited was the Moodle (Modular Object-Oriented Dynamic Learning Environment), a Virtual Learning Environment created in 1999 and multiplatform developed collaboratively by a virtual community spread around the world. According to Felix *et al.* (2018), Moodle has been the most widely used open-source virtual learning environment for distance education worldwide. Other important considerations of the tools: 5 different MOOC's are used, which represents a great acceptance of this platform in the distance learning and teaching process; Wang and Wang (2019) and Oeda and Hashimoto (2017) are not specific in the identification; Imran *et al.* (2019) use two different MOOC for this purpose.

Regarding the second dimension, 15 tools were used to extract or analyze the stored data from the teaching platform, as shown in Table 7. The highlight for the tool Waikato Environment for Knowledge Analysis (Weka), developed by the University of Waikato,

Table 7
Tools used in articles to extract or analyze data

Tools	Number of articles	References
Weka	8	Mishra and Mishra (2018); Cobos and Olmos (2019); Kostopoulos *et al.* (2019b); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Dos Santos and Falcão (2017); Rabelo *et al.* (2017)
R Studio	3	Macedo *et al.* (2019); Islam *et al.* (2019); Ramos *et al.* (2018)
MatLab	2	Mishra and Mishra (2018); Tomasevic *et al.* (2020)
JCLAL (Java Class Library for Active Learning)	1	Kostopoulos *et al.* (2018b)
DB Extractor	1	Brito *et al.* (2019)
Google Charts	1	Brito *et al.* (2019)
ERP Oracle	1	Ortigosa *et al.* (2019)
SPMF	1	De Almeida Neto and Castro (2015)
MySQL WorkBench	1	Macedo *et al.* (2019)
PGAdmin III	1	Macedo *et al.* (2019)
Microsoft Excel	1	Macedo *et al.* (2019)
MapReduce	1	Islam *et al.* (2019)
Hadoop	1	Islam *et al.* (2019)
SPSS	1	Santos *et al.* (2015)
R Project	1	Ramos *et al.* (2018)

Table 8
Tools developed

Tools	EDM and LA techniques	References
SPA (Sistema de Prediccioen de Abandono)	Classification	Ortigosa *et al.* (2019)
Dropout Risk Report	Not Specified	Brito *et al.* (2019)
EdX-MAS+ (Model Analyzer System Plus)	Classification, Prediction, Clustering	Cobos and Olmos (2019)
LOGIT Act	Prediction	De La Peña *et al.* (2018)
Prediction Tool	Classification, Regression	Kostopoulos *et al.* (2018a)
REA 2.0	Classification	Santos *et al.* (2015)
Undisclosed names	Classification, Prediction, Clustering	Chen and Zhang (2017); Whitehill *et al.* (2017); Dos Santos and Falcão (2017); Queiroga *et al.* (2017)

in New Zealand, used in 8 works for this purpose, an essential ally in combating dropout in distance education. An article can use more than one tool, as in Macedo *et al.* (2019) and Brito *et al.* (2019).

In the third dimension, 10 new products stand out, as shown in Table 8. The tool presented in Cobos and Olmos (2019), EdX-MAS+, was tested in 7 MOOC's and used 3 different techniques to analyze the data contained in the distance learning platforms: classification, prediction, and clustering. Brito *et al.* (2019) presented the Dropout Risk Report tool capable of demonstrating a list of students at risk of dropout through graphical reports. Chen and Zhang (2017), Whitehill *et al.* (2017), Dos Santos and Falcão (2017), and Queiroga *et al.* (2017) presented the results obtained by the tools developed, however, they did not disclose the names of their innovative systems.

RQ4. *What were the algorithms used to deal with dropout in distance education?*

There are several mentioned algorithms, each with its application purpose. When it comes to the prediction, detection, diagnosis, or monitoring of students able to dropout, many of the data mining has been used in educational environments. The data in Table 9 show 53 algorithms found in the studied articles, with emphasis on Random Forest, SVM, Logistic Regression and Naive Bayes, used in a number of 18, 14, 12 and 10 articles respectively. The Random Forest algorithm is well accepted, as it is simple to implement, and the working methodology is based on the creation of decision trees to obtain accurate results in the classification, prediction and regression tasks.

Regarding the 53 cited algorithms, many authors used one or more algorithms to obtain the results, according to the work proposal. Kostopoulos *et al.* (2018a), for example, used 12 algorithms in order to predict student performance in final exams. De Almeida Neto and Castro (2015), Macedo *et al.* (2019), Chen and Zhang (2017), Whitehill *et al.* (2017), Niu *et al.* (2018) and Santos *et al.* (2015) used only 1 algorithm, and Brito *et al.* (2019) and Imran *et al.* (2019) did not mention the algorithms used.

Table 9
Algorithms used in each primary article

Algorithms	Number of articles	References
Bayes Net	4	Kostopoulos *et al.* (2018b); Silva *et al.* (2015); Santos *et al.* (2016); Queiroga *et al.* (2017)
J48 Decision Tree	8	Kostopoulos *et al.* (2018b); Heidrich *et al.* (2018); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Rabelo *et al.* (2017); Queiroga *et al.* (2017); Queiroga *et al.* (2019)
Logistic Regression	12	Kostopoulos *et al.* (2018b); Liang *et al.* (2016); De La Peña *et al.* (2018); Wang and Wang (2019); Waheed *et al.* (2020); Tomasevic *et al.* (2020); Wang *et al.* (2017); Kang andWang (2018); Borrella *et al.* (2019); Ramos *et al.* (2017); Ramos *et al.* (2018); Queiroga *et al.* (2019)
Decision Tree	3	Wang and Wang (2019); Tomasevic *et al.* (2020); Wu *et al.* (2019)
Multilayer Perceptrons	4	Kostopoulos *et al.* (2018b); Santos *et al.* (2016); Queiroga *et al.* (2017); Queiroga *et al.* (2019)
Naïve Bayes	10	Kostopoulos *et al.* (2018b); Cobos and Olmos (2019); Isidro *et al.* (2018); Heidrich *et al.* (2018); Tomasevic *et al.* (2020); Kang and Wang (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2018a); Silva *et al.* (2015); Queiroga *et al.* (2019)
Gaussian Naive Bayes	3	Wang *et al.* (2017); Wu *et al.* (2019); Kostopoulos *et al.* (2019a)
Random Forest	18	Kostopoulos *et al.* (2018b); Brandão *et al.* (2019); Ortigosa *et al.* (2019); Liang *et al.* (2016); Mishra and Mishra (2018); Cobos and Olmos (2019); Wang and Wang (2019); Chen and Zhang (2017); Wang *et al.* (2017)Kang and Wang (2018); Wu *et al.* (2019); Borrella *et al.* (2019); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2019a); Santos *et al.* (2016); Dos Santos and Falcão (2017); Queiroga *et al.* (2017); Queiroga *et al.* (2019)
Sequencial Minimal Optimization (SMO)	3	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a)
Linear Regression	5	Tomasevic *et al.* (2020); Kang and Wang (2018); Wu *et al.* (2019); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a)
Bayesian Regression	1	Tomasevic *et al.* (2020)
Classification and Regression Trees (CART)	4	Brandão *et al.* (2019); Mishra and Mishra (2018); Isidro *et al.* (2018); Silva *et al.* (2015)
AdaBoost	7	Brandão *et al.* (2019); Isidro *et al.* (2018);Wang *et al.* (2017); Wu *et al.* (2019); Kostopoulos *et al.* (2018a); Santos *et al.* (2016); Queiroga *et al.* (2019)
LogitBoost	1	Kostopoulos *et al.* (2018a)
Simple Logistic	1	Queiroga *et al.* (2017)
K-nearest Neighbor (KNN)	7	Cobos and Olmos (2019); Tomasevic *et al.* (2020); Kang and Wang (2018); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2019a); Ramos *et al.* (2017); Ramos *et al.* (2018)
K-means	3	Brandão *et al.* (2019); Oeda and Hashimoto (2017); Islam *et al.* (2019)
K-means++	1	Oeda and Hashimoto (2017)
K-mediods	1	Oeda and Hashimoto (2017)
C4.5	3	Mishra and Mishra (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2018a)
C5.0	1	Ortigosa *et al.* (2019)

Continued on next page

Table 9 – continued from previous page

Algorithms	Number of articles	References
Rotation Forest	1	Kostopoulos *et al.* (2018a)
Apriori Inverse	1	De Almeida Neto and Castro (2015)
Support Vector Machine (SVM)	14	Liang *et al.* (2016); Cobos and Olmos (2019); De La Peña *et al.* (2018); Isidro *et al.* (2018); Wang and Wang (2019); Waheed *et al.* (2020); Heidrich *et al.* (2018); Tomasevic *et al.* (2020);Wang *et al.* (2017); Kang andWang (2018);Wu *et al.* (2019); Santos *et al.* (2016); Ramos *et al.* (2017); Ramos *et al.* (2018)
Radial Basis Function	1	Kostopoulos *et al.* (2018a)
Support Vector Machine (SVM) + RBF Kernel	2	Wang *et al.* (2017); Wu *et al.* (2019)
Gradient Boosting Decision Tree (GBDT)	5	Liang *et al.* (2016); Cobos and Olmos (2019); Wang and Wang (2019); Wang *et al.* (2017); Wu *et al.* (2019)
Extreme Gradient Boosting (XGBoost)	1	Niu *et al.* (2018)
Long Short Term Memory (LSTM)	2	Isidro *et al.* (2018); Wu *et al.* (2019)
Boosted Logistic Regression	1	Cobos and Olmos (2019)
Stochastic Gradient Boosting	1	Cobos and Olmos (2019)
Neuronal Network (NN)	1	Cobos and Olmos (2019)
Convolutional Neural Network (CNN)	1	Wu *et al.* (2019)
Bayesian Generalized Linear Model (BGLN)	1	Cobos and Olmos (2019)
Feed-Forward Neural Network	1	De La Peña *et al.* (2018)
Probabilistic Ensemble Simplified Fuzzy (PESFAM)	1	De La Peña *et al.* (2018)
System for Educational Data Mining	1	De La Peña *et al.* (2018)
Fuzzy C-means	1	Macedo *et al.* (2019)
Dynamic Time Warping	1	Oeda and Hashimoto (2017)
Artificial Neural Network (ANN)	4	Heidrich *et al.* (2018); Tomasevic *et al.* (2020); Ramos *et al.* (2017); Ramos *et al.* (2018)
Deep Artificial Neural Network (Deep ANN)	2	Waheed *et al.* (2020); Imran *et al.* (2019)
ID3	1	Rabelo *et al.* (2017)
Inter-Quartile Range	1	Islam *et al.* (2019)

Continued on next page

Table 9 – continued from previous page

Algorithms	Number of articles	References
L2-Regularized Logistic Regression	1	Whitehill *et al.* (2017)
M5 Rules	2	Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a)
M5 Model Tree	2	Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a)
Reduced Error Pruning Tree (REPTree)	1	Kostopoulos *et al.* (2018a)
BFTree	1	Silva *et al.* (2015)
3-Nearest Neighbor	1	Kostopoulos *et al.* (2018a)
Extra Tree Classifier (EXTRA)	1	Kostopoulos *et al.* (2019a)
JRip	1	Santos *et al.* (2016)
IBK		Santos *et al.* (2016)
Expectation Maximization (EM)	1	Dos Santos and Falcão (2017)
Not Specified	2	Brito *et al.* (2019); Imran *et al.* (2019)

For innovation, new algorithms were developed by some authors of the primary articles studied in an attempt to contribute to the identification and mitigation of dropout in distance education. According to Table 10, 10 EDM and LA algorithms were developed. Kostopoulos *et al.* (2018a), for example, used 12 algorithms and proposed the development of a new algorithm, although not named, for the dropout mitigation process in distance education. The MSSRA algorithm, proposed by Kostopoulos *et al.* (2019b), uses the techniques of classification and prediction and regression to assist in the identification of possible dropouts from the courses offered.

Table 10
Algorithms developed

Algorithms	EDM and LA techniques	References
PCA (Principal Component Analysis)	Prediction	Mishra and Mishra (2018)
E-LSTM	Classification, Prediction	Wang and Wang (2019)
ConRec Network	Classification, Prediction	Wang *et al.* (2017)
CLMS-Net	Classification, Prediction	Wu *et al.* 2019)
MSSRA	Classification, Prediction, Regression	Kostopoulos *et al.* (2019b)
Extra	Classification, Prediction	Kostopoulos *et al.* (2019a)
GBC	Classification, Prediction	Kostopoulos *et al.* (2019a)
CRISP-EDM	Classification, Prediction	Ramos *et al.* (2017)
Undisclosed names	Classification, Prediction, Regression	Kostopoulos *et al.* (2018a); Queiroga *et al.* (2019)

There is no way to distinguish between the algorithms developed to the one that had the best performance and result in the application. The fact is that, in the last five years, at least 1 algorithm is proposed each year to mine educational data and assist in learning analysis, which demonstrates once again the scientific relevance of research and the search for new techniques and solutions for mitigating dropout.

RQ5. *What were the attributes used in dropout studies in distance education?*

There are several computational tools used in distance education that collect academic data from registered users, be it the teacher, tutor, student, or manager. These data are stored in columns in the database of virtual platforms called attributes, which have each one some specific characteristic to be studied or explored.

Based on this definition and to answer this question, we categorized the responses into 4 data characteristics, as shown in Fig. 3:

(i) Demographic.
(ii) Behavioral.
(iii) Interaction.
(iv) Performance.

As a methodology, the attributes were collected from each article. We identified 36 articles with analysis of interaction data, 28 with performance data, 21 with behavioral data, and 11 with demographic data. It is worth mentioning that an article may have used more than one type of data category, such as, for example, demographic + interaction, interaction + performance, behavior + interaction, or demographic + interaction + performance.

Demographic data are attributes for the definition of users of educational platforms stored as administrative records, pre or post-entry in the courses offered. Table 11 shows the list of 10 attributes that characterize these data, the number of articles they used, and the references. Among the 10 attributes reported, Gender is the most used (8 articles), followed by Age (6 articles). Some institutions consider the attribute Number of children important for the course, as a research subsidy and organization of available time for studies, reported in 3 articles. Isidro *et al.* (2018), Waheed *et al.* (2020), and Niu *et al.*

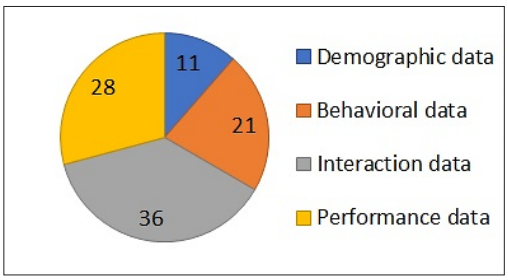

Fig 3. Number of articles by data types

Table 11
List of articles by demographic attributes

Attributes	Number of articles	References
Gender	8	Kostopoulos *et al.* (2018b); Ortigosa *et al.* (2019); Tomasevic *et al.* (2020); Kang and Wang (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a)
Age	6	Kostopoulos *et al.* (2018b); Ortigosa *et al.* (2019); Tomasevic *et al.* (2020); Kang and Wang (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2018a)
Number of children	3	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2018a)
Marital status	2	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2018a)
Computer knowledge	2	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2018a)
Computer use	2	Kostopoulos *et al.* (2018b); Kostopoulos *et al.* (2018a)
Region / City	2	Tomasevic *et al.* (2020); Kang and Wang (2018)
Type of occupation	2	Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2018a)
Nationality	1	Kostopoulos *et al.* (2015)
Work schedule	1	Kostopoulos *et al.* (2018b)

(2018) reported that they used demographic data for the analysis, however, they were not specific in defining which attributes.

Behavioral data represent the mechanisms used to access the educational platform for the user to interact. Table 12 presents 3 attributes extracted from the log records of the teaching platforms, the number of articles, and the references. According to the extracted data, 19 articles studied the attribute access to the course, since the number of times the user logs into the platform, date and time are important to check their willingness, availability, and motivation for learning. About the login duration attribute,

Table 12
List of articles by behavioral data attributes

Attributes	Number of articles	References
Course access	19	Brito *et al.* (2019); De Almeida Neto and Castro (2015); Liang *et al.* (2016); Mishra and Mishra (2018); Heidrich *et al.* (2018); Tomasevic *et al.* (2020); Chen and Zhang (2017); Wang *et al.* (2017); Imran *et al.* (2019); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2019a); Santos *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Rabelo *et al.* (2017); Ramos *et al.* (2018)
Login duration	7	Wang and Wang (2019); Kang and Wang (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Kostopoulos *et al.* (2019a); Ramos *et al.* (2017); Ramos *et al.* (2018)
Web browsers	5	Liang *et al.* (2016); Mishra and Mishra (2018); Wang and Wang (2019); Wang *et al.* (2017); Kostopoulos *et al.* (2015)

7 articles used this data to verify how long the user remains on the teaching platform, from their entry to the time they leave the environment. Finally, 5 articles cited data from web browsers for accessing the teaching platform as an important record, capable of helping to identify possible flaws in the user's interaction process with the learning environment.

Regarding the interaction data, which represent the user's involvement with the online educational platform's tools after login, the primary articles cited 10 attributes. Table 13 shows 24 articles that used discussion forum attribute as an object of study, since the non-participation of students in this activity or what they write represents signs of discouragement or lack of motivation in the course, and consequently, drop-out. Then, 14 articles cited the video attribute, which represents the user's behavior when watching a video lesson, such as, for example, how many times he pauses, plays, and starts the exhibition. This interaction process indicates a lot about the student's understanding and difficulty in learning. Another attribute mentioned and quite relevant

Table 13
List of articles by interaction data attributes

Attributes	Number of articles	References
Discussion Forum	24	Brito *et al.* (2019); Brandão *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); Liang *et al.* (2016); Mishra and Mishra (2018); Cobos and Olmos (2019); Macedo *et al.* (2019); Isidro *et al.* (2018); Wang and Wang (2019); Heidrich *et al.* (2018); Chen and Zhang (2017); Wang *et al.* (2017); Imran *et al.* (2019); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2019a); Santos *et al.* (2015); Silva *et al.* (2015); Santos *et al.* (2016); Dos Santos and Falcão (2017); Rabelo *et al.* (2017); Ramos *et al.* (2018)
Videos	14	Liang *et al.* (2016); Mishra and Mishra (2018); Cobos and Olmos (2019); Macedo *et al.* (2019); Wang and Wang (2019); Chen and Zhang (2017); Whitehill *et al.* (2017); Wang *et al.* (2017); Imran *et al.* (2019); Niu *et al.* (2018); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Santos *et al.* (2016); Dos Santos and Falcão (2017)
Messenger	11	Kostopoulos *et al.* (2018b); Brandão *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); Macedo *et al.* (2019); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a); Ramos *et al.* (2017); Ramos *et al.* (2018)
Materials made available	7	Brito *et al.* (2019); Macedo *et al.* (2019); Wang and Wang (2019); Heidrich *et al.* (2018); Santos *et al.* (2015); Santos *et al.* (2016); Dos Santos and Falcão (2017)
Wiki	4	Mishra and Mishra (2018); Wang and Wang (2019); Kostopoulos *et al.* (2015); Santos *et al.* (2016)
Chat	3	Brandão *et al.* (2019); Heidrich *et al.* (2018); Santos *et al.* (2016)
Pages visited	3	Liang *et al.* (2016); Macedo *et al.* (2019); Chen and Zhang (2017)
Web-conferences	2	Ramos *et al.* (2017); Ramos *et al.* (2018)
News	1	Kostopoulos *et al.* (2019b)
Email	1	Borrela *et al.* (2019)

is the messenger, which allows asynchronous communication between teacher, tutor, and student. The amount of interaction in this attribute represents the student's motivation, doubts, and persistence in the course. Other attributes presented in the teaching and learning process via virtual learning environments are: materials made available, which store the visit data in the course files; wiki, which stores concept and information construction data; chat, responsible for storing synchronous communication data between users; visited pages, which stores users' navigation data on the course pages; web-conferences, which store the participation and interaction of users in classes reproduced in real-time; news, responsible for storing the user's participation in the course information notes; and e-mail, which stores asynchronous communication between participants.

The articles of Oeda and Hashimoto (2017), Waheed *et al.* (2020), Tomasevic *et al.* (2020), Islam *et al.* (2019), Queiroga *et al.* (2017) and Queiroga *et al.* (2019) reported that they used interaction data for the analysis. However, they were not specific in defining which attributes were used.

As a final analysis of the attributes used in the primary articles, the performance data from the 3 attributes mentioned represent the degree of learning and grades obtained in the process. The data in Table 14 show the attributes that represent the students' level of learning. The Tasks attribute was used for analysis in 12 articles and stores the grade obtained by students in exercises and evaluative activities prepared by teachers in the virtual environment. The Questionnaires attribute, used in 11 articles, stores the grade obtained from the objective assessments carried out on the online digital platform, and finally, the Assessments attribute, used for analysis in 4 articles, stores the grade obtained in the written assessment. Isidro *et al.* (2018), Waheed *et al.* (2020), Tomasevic *et al.* (2020), Islam *et al.* (2019), Kang and Wang (2018), Kostopoulos *et al.* (2019b), Kostopoulos *et al.* (2018a) and Kostopoulos *et al.* (2019a) reported that they used performance data, such as grades obtained, for the analysis, without specifying which activities were developed. Consequently, not possible to identify the attributes used.

Table 14
List of articles by performance data attributes

Attributes	Number of articles	References
Tasks	12	Brito *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); Heidrich *et al.* (2018); Wang *et al.* (2017); Kostopoulos *et al.* (2015); Wu *et al.* (2019); Santos *et al.* (2015); Santos *et al.* (2016); Dos Santos and Falcão (2017); 2017; Ramos *et al.* (2018)
Questionnaires	11	Brito *et al.* (2019); Brandão *et al.* (2019); Ortigosa *et al.* (2019); Mishra and Mishra (2018); Macedo *et al.* (2019); Chen and Zhang (2017); Whitehill *et al.* (2017); Niu *et al.* (2018); Santos *et al.* (2016); Dos Santos and Falcão (2017); Rabelo *et al.* (2017)
Assessments	4	Kostopoulos *et al.* (2018b); De La Peña *et al.* (2018); Niu *et al.* (2018); Kostopoulos *et al.* (2015)

According to the attributes presented in the works, we can observe the quality of characteristics available, whether demographic, behavioral, interaction or performance. This set of aspects is considered essential to mitigate dropout, as they bring expressive meanings that indicate the possibility to understand how and why dropout occurred. According to Da Costa and Gouveia (2018), no single factor can cause a student to abandon an online course. Therefore, researchers recognize that it is the interaction of several factors that eventually lead a student to complete or not a course.

RQ6. *What is the level of education of the target audience in studies on dropout in distance education?*

The concern with student dropout reaches several levels of education, from high school to higher education. For this reason, educational institutions have adopted strategies that permeate between prediction, detection, diagnosis, or monitoring. The studies of the 38 primary articles brought relevant insight on this issue.

According to Table 15, which shows the level of education, the number of articles and references, 19 studied dropout in distance education in undergraduate courses, as students more easily accept teaching platforms as learning tools and understand that minimal knowledge technology and equipment with internet access is sufficient for participation in distance learning. In technical high school, 2 articles investigated dropout in distance education. Only Borrella *et al.* (2019) researched graduate studies, and 16 articles did not inform where they conducted the research; however, they acted in an attempt to mitigate dropout in distance education. Therefore, the research shows that the high dropout rate from high school to graduate school in this type of education is a concern of institutions.

Table 15
List of articles by the level of education of the target audience

Education Level	Number of articles	References
Higher Education – Postgraduate	1	Borrella *et al.* (2019)
Higher Education – Graduation	19	Kostopoulos *et al.* (2018b); Brandão *et al.* (2019); Ortigosa *et al.* (2019); De Almeida Neto and Castro (2015); De La Peña *et al.* (2018); Macedo *et al.* (2019); Isidro *et al.* (2018); Heidrich *et al.* (2018); Chen and Zhang (2017); Kang and Wang (2018); Kostopoulos *et al.* (2015); Kostopoulos *et al.* (2019b); Kostopoulos *et al.* (2018a); Kostopoulos *et al.* (2019a); Santos *et al.* (2015); Santos *et al.* (2016); Ramos *et al.* (2017); Rabelo *et al.* (2017); Ramos *et al.* (2018)
Technical High School	2	Silva *et al.* (2015); Queiroga *et al.* (2017)
Not Specified	16	Brito *et al.* (2019); Liang *et al.* (2016); Mishra and Mishra (2018); Cobos and Olmos (2019); Wang and Wang (2019); Oeda and Hashimoto (2017); Waheed *et al.* (2020); Tomasevic *et al.* (2020); Islam *et al.* (2019); Whitehill *et al.* (2017); Wang *et al.* (2017); Imran *et al.* (2019); Niu *et al.* (2018); Wu *et al.* (2019); Dos Santos and Falcão (2017); Queiroga *et al.* (2019)

RQ7. *Was an Active Methodology used to mitigate dropout in distance education?*

With the analysis of the 38 primary articles studied, we can see that none of the studies used directly Active Methodology to address dropout and enhance the permanence in distance education after the prediction, detection, diagnosis or monitoring of students from high school to postgraduate with techniques, educational data mining algorithms, and applications and learning analysis in virtual learning environments.

However, some studies report using the methodology to assist teachers in the teaching and learning process, and consequently, in dealing with dropout in distance education, in situations where the identification does not occur through EDM and LA. It is worth mentioning that these articles were not returned according to search strings. In this regard, they consider developing an Educational Recommendation System for teachers, tutors, and students that provides the use of the Active Methodology for those identified with low performance in the disciplines would help mitigate dropout, as it encourages interaction between users and the virtual environment, and collaborates with collaborative and pedagogical practices.

In the same vein Kostopoulos *et al.* (2019b) state that an interesting aspect is the implementation of semi-supervised and active learning techniques in the educational field, to predict student performance and dropout rates in educational institutions. The effectiveness and dynamics of these approaches lead to even more accurate and robust predictive models for the discovery of knowledge in educational data.

Therefore, in this perspective, using the Active Methodology in virtual learning environments would favor a more effective relationship between the actors and would collaborate to reduce the dropout rate of courses offered in distance education. It would also assist teachers and tutors in teaching practices and make them more proactive, capable of monitoring student performance with preventive measures and actions in search of positive results, and mitigating the risks of dropping out. According to Rigo *et al.* (2014), the analysis of the set of existing information would provide clues to seek methodologies to minimize dropout and enhance the permanence of students in distance education.

5. Conclusion and Future Research

This work presented a Systematic Mapping of Literature about the world scenario of research carried out on the use of the Active Methodology to mitigate dropout and motivate the permanence of students identified by the techniques of Educational Data Mining and Learning Analysis in distance learning courses. The systematic mapping of 38 articles demonstrated the growing interest in the issue of dropout in distance education, a reason for research in several countries around the world.

The high dropout rate of students in courses offered in the distance learning modality worries the managers and teachers of educational institutions, who are looking

for alternatives to identify situations that motivate students to stay in their courses. Among the options, the use of Active Methodology has been inserted in online distance learning platforms, as they present factors for improvement in student learning and promote interaction, communication, the development of critical sense, and self-learning.

Through the analysis performed, we found that some works use the Active Methodology to assist teachers in the teaching and learning process, in an attempt to reduce the dropout rate of the courses and enhance their permanence. However, none of the studies did consider integrating Active Methodologies with the techniques of EDM and LA, in order to reduce the risks of dropout.

The use of EDM and LA started from the conception that both use the same techniques or methods of application and have similar definitions and objectives in the teaching process. These characteristics for the analysis of educational data and the provision of information can support decision making, collaborate to identify the risks of dropout, and enhance the permanence of students in distance education, the main focus of educational institutions. Also, helping the teacher to monitor the student's performance during the learning process, a factor that mitigates the dropout problem.

The mapping results demonstrated that dropout is the focus of several studies, and according to the articles studied, there was little evidence of using an Active Methodology for this purpose after the identification by EDM and LA techniques. Nevertheless, some works consider the possibility of adding Active Methodologies to the techniques of data mining and learning analysis, to contribute to dropout mitigating and permanence increasing of students (Chandrasekaran *et al.,* 2016; Leite and Ramos, 2017; Lima and Siebra, 2017).

The mining and learning analysis is widely used, which we have identified: (i) 5 techniques, with emphasis on classification and prediction; (ii) 15 existing tools widely used and 10 new proposals developed; (iii) 53 different algorithms used and 10 new ones developed; (iv) the attributes that store records of demographic data, behavior, interaction, and performance are frequently used, with emphasis on the interaction data present in 94% of the studies; (v) 50% of the articles focused on research at higher education, undergraduate level.

As future work we intend to pursue (i) the identification of Active Methods, among the several existing and mentioned in this work, that can be applied to the context of mitigating dropout and enhancing permanence in distance education; (ii) the development of a Recommendation System that uses an Active Methodology to improve performance and mitigate the risks of failure and dropout of the student in the course.

References

Baker, R., Isotani, S., Carvalho, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19, 3–13. DOI: 10.5753/rbie.2011.19.02.03.

Baker, R.S.J.D., Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1, 3–16. DOI: 10.5281/zenodo.3554657.

Borrella, I., Caballero-Caballero, S., Ponce-Cueto, E. (2019). Predict and intervene: Addressing the dropout problem in a MOOC-based program. In: *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S 2019)*, pp. 1–9. DOI: 10.1145/3330430.3333634.

Brandão, I.V., Da Costa, J.P.C., Santos, G.A., Praciano, B.J., Junior, F.C., Rafael, R.T. (2019). Classification and predictive analysis of educational data to improve the quality of distance learning courses. In: *4th Workshop on Communication Networks and Power Systems (WCNPS 2019)*. DOI: 10.1109/WCNPS.2019.8896312.

Brito, M., Medeiros, F., Bezerra, E.P. (2019). An infographics-based tool for monitoring dropout risk on distance learning in higher education. In: *18th International Conference on Information Technology Based Higher Education and Training (ITHET 2019)*, IEEE. pp. 1–7. DOI: 10.1109/ITHET46829.2019.8937361.

Cambruzzi,W., Rigo, S.J., Barbosa, J.L. (2015). Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. *Journal of Universal Computer Science*, 21, 23–47.

Chandrasekaran, S., Badwal, P., Thirunavukkarasu, G., Littlefair, G. (2016). Collaborative learning experience of students in distance education. In: *International Symposium on Project Approaches in Engineering Education*, 6, 90–99.

Chen, Y., Zhang, M. (2017). MOOC student dropout: pattern and prevention. In: *ACM Turing 50th Celebration Conference (ACM TUR-C 2017)*, pp. 1–6. DOI: 10.1145/3063955.3063959.

Cobos, R., Olmos, L. (2019). A Learning Analytics Tool for Predictive Modeling of Dropout and Certificate Acquisition on MOOCs for Professional Learning. In: *IEEE International Conference on Industrial Engineering and Engineering Management*, IEEE. pp. 1533–1537. DOI: 10.1109/IEEM.2018.8607541.

Da Costa, O.S., Gouveia, L.B. (2018). Dropout in distance learning: A reference model for an integrated alert system. In: *Euro American Conference on Telematics and Information Systems (EATIS 2018)*, pp. 1–5. DOI: 10.1145/3293614.3293648.

De Almeida Neto, F.A., Castro, A. (2015). Elicited and mined rules for dropout prevention in online courses. In: *IEEE Frontiers in Education Conference (FIE 2015)*. DOI: 10.1109/FIE.2015.7344048.

De La Peña, D., Lara, J.A., Lizcano, D., Martínez, M.A., Burgos, C., Campanario, M.L. (2018). Mining activity grades to model students' performance. In: *2017 International Conference on Engineering and MIS (ICEMIS 2017)*, pp. 1–6. DOI: 10.1109/ICEMIS.2017.8272963.

Dos Santos, D.C.V.B., Falcão, T.P. (2017). Acompanhamento de alunos em ambientes virtuais de aprendizagem baseado em sistemas tutores inteligentes. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, pp. 1267–1276. DOI: 10.5753/cbie.sbie.2017.1267.

Felix, I., Ambrósio, A.P., Lima, P.d.S., Brancher, J.D. (2018). Data Mining for Student Outcome Prediction on Moodle: a systematic mapping. In: *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*, pp. 1393–1402. DOI: 10.5753/cbie.sbie.2018.1393.

Gokhale, A.A. (1995). Collaborative Learning Enhances Critical Thinking. *Journal of Technology Education*, 7, 22–30. DOI: 10.21061/jte.v7i1.a.2.

Guo, R., Li, L., Han, M. (2018). On-demand virtual lectures: Promoting active learning in distance learning. In: *2nd International Conference on E-Education, E-Business and E-Technology (ICEBT 2018)*, pp. 1–5. DOI: 10.1145/3241748.3241757.

Heidrich, L., Victória Barbosa, J.L., Cambruzzi, W., Rigo, S.J., Martins, M.G., dos Santos, R.B.S. (2018). Diagnosis of learner dropout based on learning styles for online distance learning. *Telematics and Informatics*, 35, 1593–1606. DOI: 10.1016/j.tele.2018.04.007.

Imran, A.S., Dalipi, F., Kastrati, Z. (2019). Predicting student dropout in a MOOC: An evaluation of a deep neural network model. In: *5th International Conference on Computing and Artificial Intelligence (ICCAI 2019)*, pp. 190–195. DOI: 10.1145/3330482.3330514.

Isidro, C., Carro, R.M., Ortigosa, A. (2018). Dropout detection in MOOCs: An exploratory analysis. In: *2018 International Symposium on Computers in Education (SIIE 2018)*, pp. 1–6. DOI: 10.1109/SIIE.2018.8586748.

Islam, O., Siddiqui, M., Aljohani, N.R. (2019). Identifying online profiles of distance learning students using data mining techniques. In: *3rd International Conference on Digital Technology in Education (ICDTE 2019)*, pp. 115–120. DOI: 10.1145/3369199.3369249.

Kang, K., Wang, S. (2018). Analyze and predict student dropout from online programs. In: *2nd International Conference on Compute and Data Analysis (ICCDA 2018)*, pp. 6–12. DOI: 10.1145/3193077.3193090.

Kostopoulos, G., Karlos, S., Kotsiantis, S. (2019a). Multiview Learning for Early Prognosis of Academic Performance: A Case Study. *IEEE Transactions on Learning Technologies*, 12, 212–224. DOI: 10.1109/TLT.2019.2911581.

Kostopoulos, G., Kotsiantis, S., Fazakis, N., Koutsonikos, G., Pierrakeas, C. (2019b). A Semi-Supervised Regression Algorithm for Grade Prediction of Students in Distance Learning Courses. *International Journal on Artificial Intelligence Tools*, 28, 1–19. DOI: 10.1142/S0218213019400013.

Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., Gravvanis, G.A. (2018a). Forecasting students' success in an open university. *International Journal Learning Technology*, 13, 26–43. DOI: 10.1504/IJLT.2018.091630.

Kostopoulos, G., Kotsiantis, S., Pintelas, P. (2015). Estimating student dropout in distance higher education using semi-supervised techniques. In: *19th Panhellenic Conference on Informatics (PCI 2015)*, pp. 38–43. DOI: 10.1145/2801948.2802013.

Kostopoulos, G., Kotsiantis, S., Ragos, O., Grapsa, T.N. (2018b). Early dropout prediction in distance higher education using active learning. In: *2017 8th International Conference on Information, Intelligence, Systems and Applications (IISA 2017)*, pp. 1–6. DOI: 10.1109/IISA.2017.8316424.

Leite, L.S., Ramos, M.B. (2017). *A metodologia ativa no Ambiente Virtual de Aprendizagem*. Pimenta Cultural, volume 1. pp. 85–101.

Liang, J., Yang, J., Wu, Y., Li, C., Zheng, L. (2016). Big data application in education: Dropout prediction in edx MOOCs. In: *2016 IEEE 2nd International Conference on Multimedia Big Data (BigMM 2016)*, IEEE. pp. 440–443. DOI: 10.1109/BigMM.2016.70.

Lima, E., Siebra, C. (2017). CollabEduc: Uma Ferramenta de Colaboração em Pequenos Grupos para Plataformas de Aprendizagem a Distância. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, pp. 1707–1716. DOI: 10.5753/cbie.sbie.2017.1707.

Macedo, M., Santana, C., Siqueira, H., Rodrigues, R.L., Ramos, J.L.C., Silva, J.C.S., Maciel, A. M. A., Bastos-Filho, C.J. (2019). Investigation of college dropout with the fuzzy c-means algorithm. In: *IEEE 19th International Conference on Advanced Learning Technologies (ICALT 2019)*, IEEE. pp. 187–189. DOI: 10.1109/ICALT.2019.00055.

Manhães, L.M.B., Serra da Cruz, S.M., Macário Costa, R.J., Zavaleta, J., Zimbrão, G. (2011). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE 2011)*, pp. 150–159. DOI: 10.5753/cbie.sbie.2011.25p.

Maschio, P., Vieira, M.A., Costa, N., Melo, S.D., Júnior, C.P. (2018). Um Panorama acerca da Mineração de Dados Educacionais no Brasil. In: *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*, pp. 1936–1940. DOI: 10.5753/cbie.sbie.2018.1936.

Mishra, B.B., Mishra, S. (2018). Quality Improvements in Online Education System by Using Data Mining Techniques. In: *2nd International Conference on Data Science and Business Analytics (ICDSBA 2018)*, pp. 532–536. DOI: 10.1109/ICDSBA.2018.00105.

Niu, Z., Li, W., Yan, X., Wu, N. (2018). Exploring causes for the dropout on massive open online courses. In: *ACM International Conference Proceeding Series*, pp. 47–52. DOI: 10.1145/3210713.3210727.

Oeda, S., Hashimoto, G. (2017). Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes. In: *Procedia Computer Science*, Elsevier B.V., pp. 614–621. DOI: 10.1016/j.procs.2017.08.088.

Oliveira, W. P., Bittencourt, W. J. M. (2020). A evasão na EaD: uma análise sobre os dados e relatórios, ano base 2017, apresentados pelo Inep, UAB e Abed. *Educação Pública*, 20, 3.

Ortigosa, A., Carro, R.M., Bravo-Agapito, J., Lizcano, D., Alcolea, J.J., Blanco, Ó. (2019). From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System. *IEEE Transactions on Learning Technologies*, 12, 264–277. DOI: 10.1109/TLT.2019.2911608.

Petersen, K., Vakkalanka, S., Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. In: *Information and Software Technology*, Elsevier B. V.. pp. 1–18. DOI: 10.1016/j.infsof.2015.03.007.

Queiroga, E., Cechinel, C., Aguiar, M. (2019). Uma abordagem para predição de estudantes em risco utilizando algoritmos genéticos e mineração de dados: um estudo de caso com dados de um curso técnico a distância. In: *Anais do Workshop do VIII Congresso Brasileiro de Informática na Educação (WCBIE 2019)*, pp. 119–128. DOI: 10.5753/cbie.wcbie.2019.119.

Queiroga, E., Cechinel, C., Araújo, R. (2017). Predição de estudantes com risco de evasão em cursos técnicos a distância. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, pp. 1547–1556. DOI: 10.5753/cbie.sbie.2017.1547.

Rabelo, H., Burlamaqui, A., Valentim, R., Rabelo, D.S.d.S., Medeiros, S. (2017). Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, pp 1527–1536. DOI: 10.5753/cbie.sbie.2017.1527.

Ramos, J.L.C., Gomes, A.S., Rodrigues, R., Silva, J., de Souza, F.D.F., Zambom, E.D.G., Prado, L. (2017). Um Modelo Preditivo da Evasão dos Alunos na EAD a Partir dos Construtos da Teoria da Distância Transacional. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, pp. 1227–1236. DOI: 10.5753/cbie.sbie.2017.1227.

Ramos, J.L.C., Silva, J., Prado, L., Gomes, A., Rodrigues, R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. In: *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*, pp. 1463–1472. DOI: 10.5753/cbie.sbie.2018.1463.

Rigo, S.J., Cambruzzi,W., Barbosa, J.L.V., Cazella, S.C. (2014). Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, 22, 132–146. DOI: 10.5753/rbie.2014.22.01.132.

Rodrigues, M. W., Isotani, S., Zárate, L.E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, 35, 1701–1717. DOI: 10.1016/j.tele.2018.04.015.

Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33, 135–146. DOI: 10.1016/j.eswa.2006.04.005.

Romero, C., Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3, 12–27. DOI: 10.1002/widm.1075.

Romero, C., Ventura, S., García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers and Education*, 51, 368–384. DOI: 10.1016/j.compedu.2007.05.016.

Santos, F.D., Bercht, M., Wives, L. (2015). Classificação de alunos desanimados em um AVEA: uma proposta a partir da mineração de dados educacionais. In: *Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)*, pp. 1052–1061. DOI: 10.5753/cbie.sbie.2015.1052.

Santos, R., Pitangui, C., Assis, L., Vivas, A. (2016). Uso de Séries Temporais e Seleção de Atributos em Mineração de Dados Educacionais para Previsão de Desempenho Acadêmico. In: *Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)*, pp. 1146–1155. DOI: 10.5753/cbie.sbie.2016.1146.

Siemens, G., Baker, R.S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In: *ACM International Conference Proceeding Series*, pp. 252–254. DOI: 10.1145/2330601.2330661.

Silva, F., Silva, J.D., Silva, R., Fonseca, L.C. (2015). Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. In: *Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)*, pp. 1187–1196. DOI: 10.5753/cbie.sbie.2015.1187.

Tjhin, V., Rahayu, A., Soraya, K. (2017). Evaluating the Performance of Students through Collaborative Learning. In: *10th International Conference on Human System Interaction (HSI)*, pp. 98–103. DOI: 10.1109/HSI.2017.8005006.

Thomas, L.K., Harden-Thew, K., Delahunty, J., Dean, B.A. (2016). A vision of You-topia: Personalising professional development of teaching in a diverse academic workforce. *Journal of University Teaching and Learning Practice*, 13, 1–12.

Tomasevic, N., Gvozdenovic, N., Vranes, S. (2020). An overviewand comparison of supervised data mining techniques for student exam performance prediction. *Computers and Education*, 143, 103676. DOI: 10.1016/j.compedu.2019.103676.

Waheed, H., Hassan, S.U., Aljohani, N.R., Hardman, J., Alelyani, S., Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104, 106189. DOI: 10.1016/j.chb.2019.106189.

Wang, L., Wang, H. (2019). Learning behavior analysis and dropout rate prediction based on MOOCs data. In: *Proceedings – 10th International Conference on Information Technology in Medicine and Education (ITME 2019)*, pp. 419–423. DOI: 10.1109/ITME.2019.00100.

Wang,W., Yu, H., Miao, C. (2017). Deep model for dropout prediction in MOOCs. In: *ACM International Conference Proceeding Series*, pp. 26–32. DOI: 10.1145/3126973.3126990.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., Tingley, D. (2017). MOOC dropout prediction: How to measure accuracy?. In: *L@S 2017 – Proceedings of the 4th (2017) ACM Conference on Learning at Scale*, pp. 161–164. DOI: 10.1145/3051457.3053974.

Wu, N., Zhang, L., Gao, Y., Zhang, M., Sun, X., Feng, J. (2019). CLMS-Net : Dropout Prediction in MOOCs with Deep Learning. In: *ACM Turing Celebration Conference – ACM TURC 19*, pp. 1–6. DOI: 10.1145/3321408.3322848.

T.L. de Andrade is a PhD student in the Graduate Program in Applied Computing at the University of Vale do Rio dos Sinos (UNISINOS) and teacher at the Department of Computing at the State University of Mato Grosso (UNEMAT).

S.J. Rigo is Professor/Researcher in the Applied Computing Graduate Program Program (UNISINOS); Dean of UNISINOS Polytechnic School. Head of the UNISINOS Graduate Degree Program in Applied Computer Science (2016–2017). Head of the UNISINOS Undergraduate Degree Program in Computer Science (2011–2015). Bsc in Computer Science at Pontifícia Universidade Católica do Rio Grande do Sul PUCRS (1990); MsC in Computer Science at Universidade Federal do Rio Grande do Sul UFRGS (1993); PhD in Computer Science at Universidade Federal do Rio Grande do Sul UFRGS (2008). Post-doctorate at Friedrich-Alexander Universität Erlangen-Nürnberg/Germany (2018). Professor and Researcher at UNISINOS University, since 1995. Research areas: Artificial Intelligence, Semantic Web, Natural Language Processing, Distance Education.

J.L.V. Barbosa holds a degree in Informatics (1990) and Electrical Engineering (1991) from the Catholic University of Pelotas (UCPel), obtained a specialization in Software Engineering (UCPel, 1993) and completed a master's and a doctorate in computer science at the Federal University of Rio Grande do Sul (UFRGS, 1996 and 2001). In 2011, he completed a post-doctorate at Sungkyunkwan University (SKKU, Suwon, South Korea). In 2020, he completed a post-doctorate at the University of California Irvine (UCI, Irvine, USA). He is currently Full Professor II at the University of Vale do Rio dos Sinos (Unisinos), works as a professor in the Postgraduate Program in Applied Computing (PPGCA), and the Professional Master in Electrical Engineering. Jorge served as a member of the technical committee of the 'International Conference on Communication Systems and Network Technologies (CSNT 2018, Bhopal, India)' and was a member of the 'International Advisory Committee' and 'General Chair' of the 'International Conference on Intelligent Computing and Smart Communication, (ICSC 2019, THDC-IHET, Tehri, Uttarakhand, India). They stand out as their main areas of activity: Mobile and Ubiquitous Computing, History of Contexts, Profile Management, Similarity Analysis of Historical Contexts, Prediction of Contexts, Computer Games, Ubiquitous Education, Ubiquitous Health, Ubiquitous Accessibility, Ubiquitous Agriculture and Ubiquitous Project Management.