Received: 30 September 2019 | Revised: 11 December 2019 | Accepted: 19 December 2019
DOI: 10.1002/widm.1355

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY
WILEY

Educational data mining and learning analytics: An updated survey

Cristobal Romero | **Sebastian Ventura**

Computer Sciences and Numerical Analysis, University of Cordoba, Andalusia, Spain

Correspondence
Cristobal Romero, Computer Sciences and Numerical Analysis, University of Cordoba, Andalusia, Spain.
Email: cromero@uco.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: TIN2017-83445-P

Abstract

This survey is an updated and improved version of the previous one published in 2013 in this journal with the title "data mining in education". It reviews in a comprehensible and very general way how Educational Data Mining and Learning Analytics have been applied over educational data. In the last decade, this research area has evolved enormously and a wide range of related terms are now used in the bibliography such as Academic Analytics, Institutional Analytics, Teaching Analytics, Data-Driven Education, Data-Driven Decision-Making in Education, Big Data in Education, and Educational Data Science. This paper provides the current state of the art by reviewing the main publications, the key milestones, the knowledge discovery cycle, the main educational environments, the specific tools, the free available datasets, the most used methods, the main objectives, and the future trends in this research area.

This article is categorized under:
Application Areas > Education and Learning

KEYWORDS

Educational Data Mining, Data Mining on Education, Data-Driven Decision-Making in Education, Big Data in Education, Educational Data Science

1 | INTRODUCTION

The increase of e-learning resources, instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information has created large repositories of educational data. Traditional educational institutions have used for many year information systems that store plenty of interesting information. Nowadays, web-based educational systems have been rising exponentially and they led us to store a huge amount of potential data from multiple sources with different formats and with different granularity levels (Romero & Ventura, 2017). New types of educational environments such as blended learning (BL), virtual/enhanced environments, mobile/ubiquitous learning, game learning, etc. also gather huge amount of data about students. All these systems produce huge amount of information of high educational value, but it is impossible to analyze it manually. So, tools to automatically analyze this kind of data are needed because of all this information provides a goldmine of educational data that can be explored and exploited to understand how students learn. In fact, today, one of the biggest challenges that educational institutions face is the exponential growth of educational data and the transformation of this data into new insights that can benefit students, teachers, and administrators (Baker, 2015).

Two different communities have grown around the same area with a joint interest in how educational data can be exploited to benefit education and the science of learning (Baker & Inventado, 2014):

- **Educational Data Mining (EDM)** is concerned with developing methods for exploring the unique types of data that come from educational environments (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018). It can be also defined as the application of data mining (DM) techniques to this specific type of dataset that come from educational environments to address important educational questions (Romero & Ventura, 2013).
- **Learning Analytics (LA)** can be defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Lang, Siemens, Wise, & Gasevic, 2017). There are three crucial elements involved in this definition (Siemens, 2013): data, analysis and action.

Both communities share a common interest in data-intensive approaches to educational research, and share the goal of enhancing educational practice (Siemens & Baker, 2012) (Liñán & Pérez, 2015). On the one hand, LA is focused on the educational challenge and EDM is focused on the technological challenge. LA is focused on data-driven decision-making and integrating the technical and the social/pedagogical dimensions of learning by applying known predictive models. On the other hand, EDM is generally looking for new patterns in data and developing new algorithms and/or models. Ultimately, the differences between the two communities are more based on focus, research questions, and the eventual use of models, than on the methods being used (Baker & Inventado, 2014). Regardless of the differences between the LA and EDM communities, the two areas have significant overlap both in the objectives of investigators as well as in the methods and techniques that are used in the investigation.

EDM and LA are interdisciplinary areas including but not limited to information retrieval, recommender systems, visual data analytics, domain-driven data mining, social network analysis, psychopedagogy, cognitive psychology, psychometrics, and so on. In fact, they can be drawn as the combination of three main areas (Figure 1): computer science, education, and statistics. The intersection of these three areas also forms other subareas closely related to EDM and LA such as computer-based education (CBE), data mining and machine learning, and educational statistics.

In addition to EDM and LA, there are also other related terms used in the bibliography:

- **Academic Analytics (AA) and Institutional Analytics (IA)** is concerned with the collection, analysis, and visualization of academic program activities such as courses, degree programs; research, revenue of students' fees, course evaluation, resource allocation, and management to generate institutional insight (Campbell, DeBlois, & Oblinger, 2007; Siemens and Long, 2011). So, it is focused on the political/economic challenge.
- **Teaching Analytics (TA)** refers to the analysis of teaching activities and performance data as well as the design, development, and evaluation of teaching activities (Prieto, Sharma, Dillenbourg, & Jesús, 2016). It is focused on the educational challenge from the instructors' point of view.
- **Data-Driven Education (DDE) and Data-Driven Decision-Making in Education (DDDM)** refers to systematically collect and analyze various types of educational data, to guide a range of decisions to help improve the success of students and schools (Custer, King, Atinc, Read, & Sethi, 2018; Datnow & Hubbard, 2016).
- **Big Data in Education (BDE)** refers to apply big data (basic connotation summed up in volume, variety, value, and velocity) techniques to data from educational environment (Daniel, 2019).
- **Educational Data Science (EDS)** is defined as the use of data gathered from educational environments/settings for solving educational problems (Romero & Ventura, 2017). Data science is a concept to unify statistics, data analysis, machine learning, and their related methods.

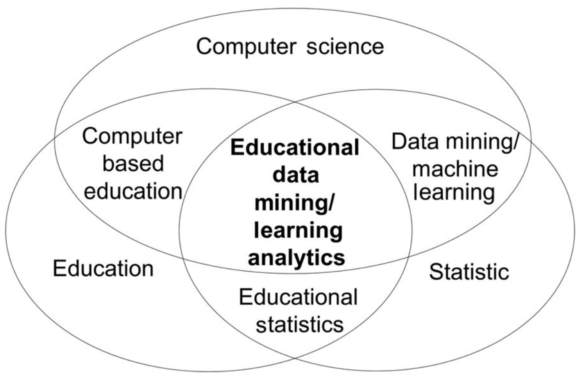


FIGURE 1 Main areas related to Educational Data Mining/Learning Analytics

This survey is an updated and improved version of the previous one published in 2013 in this journal with the title "data mining in education" (Romero & Ventura, 2013). It is needed to redo a comprehensible overview of the current state of knowledge in EDM and LA because 6 years have passed and a huge number of new papers have been published. The main changes we have observed since the previous survey are: new related terms are used in the bibliography (AA, IA, TA, DDE, DDDM, BDE, and EDS), the number of published books and papers has grown exponentially (more in LA than EDM), the interest on data related to new types of educational environments has increased (Massive Open Online Courses, MOOCs, virtual and augmented reality learning, serious games, BL, etc.), more specific tools and free datasets are available, the number of application problems or topics of interest is wider, and finally, there are new future trends.

2 | BACKGROUND

EDM and LA have emerged from two independent conferences and communities. The first Educational Data Mining Conference was in Montreal, Canada in 2008 organized by the IEDM society, and the first Learning Analytics and Knowledge Conference was in Banff, Canada in 2011 organized by the SOLAR society. There are also some other closely-related conferences (Table 1).

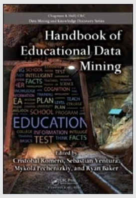
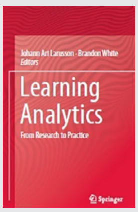
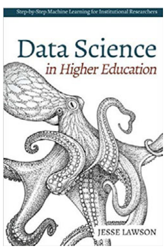
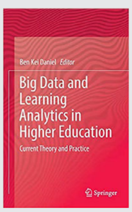
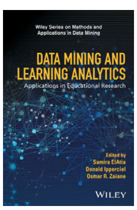
The first book about EDM/LA topics was published on 2006 and it was entitled Data Mining in E-Learning (Romero & Ventura, 2006). Since then, an increasing number of books have been published (Table 2), especially in the last years. We can see that during the first years (from 2006 to 2014) the terms Data Mining in Education and Educational Data Mining were used in the titles. Next (from 2015 to 2017) the terms Learning Analytics, Data Science, Big Data, and Data Mining were also used. And in the last years, the terms Learning Analytics is the most used in the titles. From all of them, the two most important books in the area are the *Handbook of Educational Data Mining* (Romero et al., 2010) and the *Handbook of Learning Analytics* (Lang et al., 2017). Additionally, we want to highlight that there is also an online[1] Massive Online Open Textbook (MOOT) that contains all the resources as seen on Coursera and EdX Big Data and Education courses (Baker, 2015).There are several international and prestigious journals in which most of the papers about EDM and LA have been published (Table 3). Of all of them, the two most specific journals are the *Journal of Educational Data Mining*[2] which was launched in 2009 and the *Journal of Learning Analytics*[3] which was launched in 2014. We also want to notice that there are some just born new related journals such as: *International Journal of Learning Analytics* and *Artificial Intelligence for Education (iJAI)*,[4] and *Computer-Based Learning in Context (CBLC)*[5] both launched in 2019.

Finally, in order to show the growing interest in EDM and LA during the last two decades, Figure 2 shows the number of papers or results that return a freely accessible web search engine such as *Google Schoolar* when searching the exact term "Educational Data Mining" or "Learning Analytics" in each year from 2000 to 2018. As can be seen, both numbers grow in an exponential way, showing the high interest in both topics. And although EDM has more references than LA until 2011, then LA surpass EDM. This fact can be explained by the temporal distribution of the next important events in EDM and LA history such as: the first workshop about EDM in the Association for the Advancement of Artificial Intelligence 2005 Conference, the first book related with EDM/LA (Romero & Ventura, 2006), the first Educational Data Mining

TABLE 1 Most related conferences about Educational Data Mining/Learning Analytics

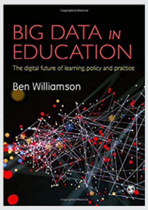
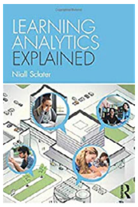
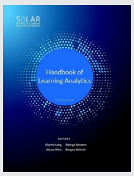
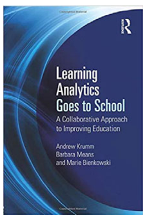
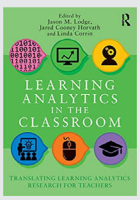
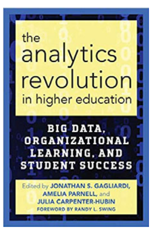
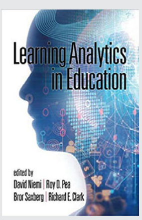
Title	Acronym	Type	1° year
International Conference on Artificial Intelligence in Education	AIED	Biannual	1982
International Conference on Intelligent Tutoring Systems	ITS	Biannual	1988
IEEE International Conference on Advanced Learning Technologies	ICALT	Annual	2000
European Conference on Technology-Enhanced Learning	EC-TEL	Annual	2006
International Conference on Educational Data Mining	EDM	Annual	2008
International Conference on User Modeling, Adaptation, and Personalization	UMAP	Annual	2009
International Conference on Learning Analytics and Knowledge	LAK	Annual	2011
Learning at Scale	L@S	Annual	2014
Learning and Students Analytics Conference	LSAC	Annual	2017

TABLE 2 Published books about Educational Data Mining/Learning Analytics

Cover	Title	Authors	Year	Editorial	Pages
	Data Mining in Education	C. Romero & *S. Ventura*	2006	Wit Press	299
	Handbook of Educational Data Mining	C. Romero, S. Ventura., M. Pechenizky, R. Baker	2010	CRC Press, Taylor & Francis Group	535
	Education Data Mining: Applications and Trends	A. Peña-Ayala	2014	Springer	468
	Learning Analytics: From research to practice.	J.A. Larusson, B. White	2014	Springer	195
	Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers	J. Lawson	2015	CreateSpace Independent Publishing Platform	226
	Big Data and Learning Analytics in Higher Education: Current Theory and Practice	B.k. Daniel	2016	Springer	272
	Data Mining and Learning Analytics: Applications in Educational Research	S. ElAtia, D. Ipperciel, O.R. Zaïane	2016	Wiley	320

(Continues)

TABLE 2 (Continued)

Cover	Title	Authors	Year	Editorial	Pages
	Big Data in Education: The digital future of learning, policy and practice	B. Williamson	2017	SAGE Publications	256
	Learning Analytics Explained	Niall Sclater	2017	Routledge	290
	Handbook of Learning Analytics	C. Lang, G. Siemens, A. Wise, D. Gašević	2017	SOLAR	356
	Learning Analytics Goes to School	A. Krumm, B. Means, M. Bienkowski	2018	Routledge	190
	Learning Analytics in the Classroom	J. Horvath, J. Lodge, L. Corrin	2018	Routledge	314
	The Analytics Revolution in Higher Education: Big Data, Organizational Learning, and Student Success	J. S. Gagliardi, A. Parnell, J. Carpenter-Hubin, R. L. Swing	2018	Stylus Publishing	252
	Learning Analytics in Education	D. Niemi, R. D. Pea, B. Saxberg, R. E. Clark	2018	Information Age Publishing	268

(Continues)

TABLE 2 (Continued)

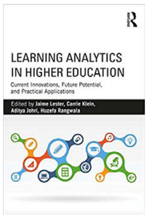
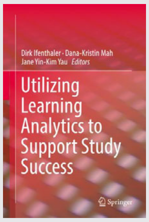
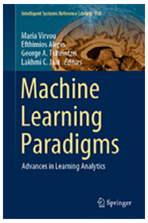
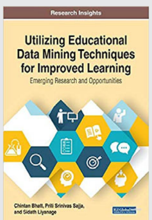
Cover	Title	Authors	Year	Editorial	Pages
	Learning Analytics in Higher Education: Current Innovations, Future Potential, and Practical Applications	J. Lester, C. Klein, A. Johri, H. Rangwala	2018	Routledge	216
	Utilizing Learning Analytics to Support Study Success	D. Ifenthaler, D. Mah, Y. J. Yau	2019	Springer	328
	Machine Learning Paradigms: Advances in Learning Analytics	M. Virvou, E. Alepis, G.A. Tsihrintzis, L.C. Jain	2019	Springer	223
	Utilizing Educational Data Mining Techniques for Improved Learning: Emerging Research and Opportunities	C. Bhatt, P.S. Sajja, S. Liyanage	2019	IGI Global	188

TABLE 3 Top related journals about Educational Data Mining/Learning Analytics

Journal title	Number of papers	Impact factor 2018	Free and open access
Journal of Learning Analytics	143	—	Yes
Computers and Education	81	5.627*	No
British Journal of Educational Technology	65	2.588**	No
Journal of Educational Data Mining	48	—	Yes
Journal of Artificial Intelligence in Education	47	—	No
IEEE Transactions on Learning Technologies	33	2.315*	No
Journal of Computer Assisted Learning	32	2.451**	No
International Journal on Technology Enhanced Learning	31	—	No
User Modeling and User-Adapted Interaction	27	3.400*	No
Internet and Higher Education	26	5.284**	No
Computer Applications in Engineering Education	26	1.435*	No

*JCR Science Edition, **JCR Social Science Edition, January 1, 2019.

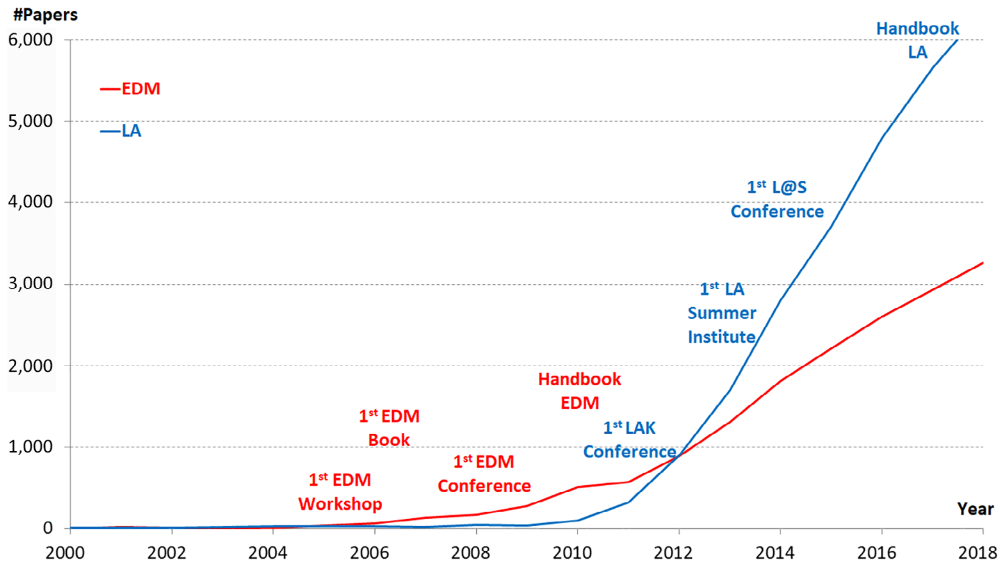


FIGURE 2 Number of papers and main events about Educational Data Mining/ Learning Analytics terms in Google Schoolar by year (January 1, 2019)

TABLE 4 Top-10 most cited papers about Educational Data Mining and Learning Analytics

Paper title	Reference	Num. Cites*	Num. cites**
Educational data mining: A survey from 1995 to 2005	(Romero & Ventura, 2007)	1,489	662
Educational data mining: A review of the state of the art	(Romero & Ventura, 2010)	1,367	631
The state of educational data mining in 2009: A review and future visions	(Baker & Yacef, 2009)	1,199	—
Penetrating the fog: Analytics in learning and education	(Siemens & Long, 2011)	1,138	-
Data mining in course management systems: Moodle case study and tutorial	(Romero, Ventura, & Salcines, 2008)	1,105	470
Learning analytics: Drivers, developments, and challenges	(Ferguson, 2012)	691	328
Learning analytics and educational data mining: Towards communication and collaboration	(Siemens & Baker, 2012)	589	224
Course signals at Purdue: Using learning analytics to increase student success	(Arnold & Pistilli, 2012)	569	206
Translating learning into numbers: A generic framework for learning analytics	(Greller & Drachsler, 2012)	547	221
Mining educational data to analyze students' performance	(Baradwaj & Pal, 2012)	543	—

*Google Schoolar, **SciVerse Scopus, January 1, 2019.

Conference in Montreal (Canada) 2009, the handbook of EDM (Romero et al., 2010), the first Learning Analytics & Knowledge in Banf (Canada) 2011, the first LA Summer Institute in Palo Alto (USA) 2013, the first Learning at Scale Conference in Atlanta (USA) 2015, and the handbook of LA (Lang et al., 2017). This greater increment in the number of papers that use the term "Learning Analytics" rather than "Educational Data Mining" produce that a bibliometric approach (Dormezil, Khoshgoftaar, & Robinson-Bryant, 2019) conclude it is more accurate to describe what appears to be two domains (i.e., Educational Data Mining and Learning Analytics) as one domain (i.e., Learning Analytics) with one prominent subset (i.e., Educational Data Mining).

Finally, the most cited papers in EDM and LA are shown in Table 4. We can see that four in 10 papers are reviews/surveys. Analyzing this important type of papers, the first and most popular review of EDM research was published by

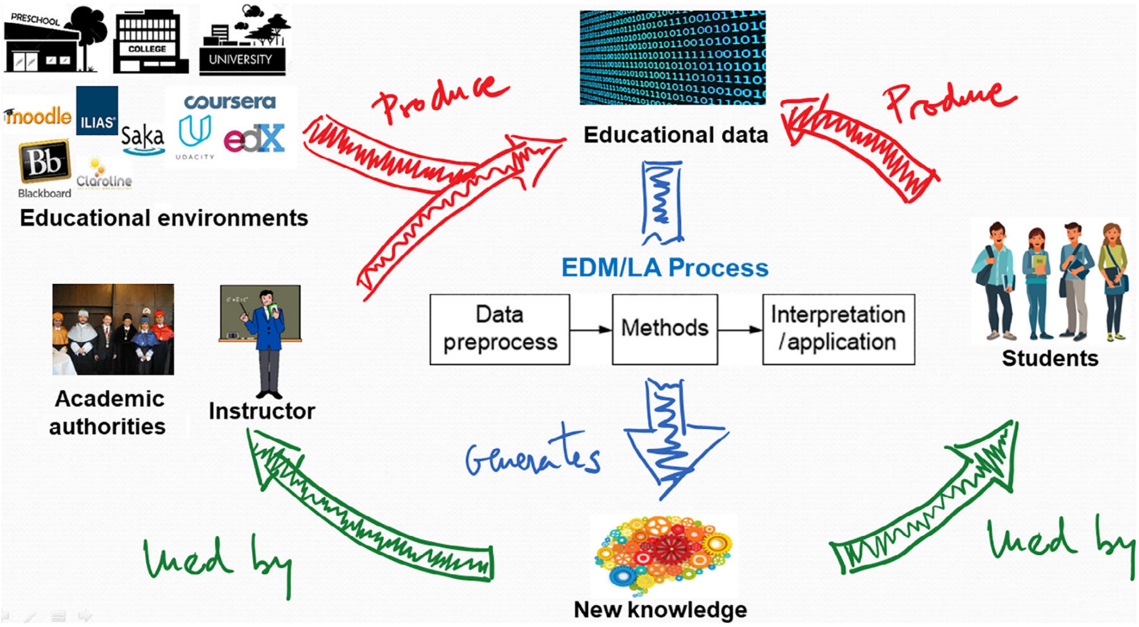


FIGURE 3 Educational Data Mining/Learning Analytics knowledge discovery cycle process

Romero and Ventura (2007), and it was followed by a more complete (Romero & Ventura, 2010) and a more comprehensible (Romero & Ventura, 2013) reviews by the same authors. Another popular review was presented in the inaugural issue of *EDM* journal (Baker & Yacef, 2009). An important report was published by the U.S. Office of Educational Technology about how to enhance teaching and learning through EDM and LA (Bienkowski, Feng, & Means, 2012). The differences between EDM and LA are deal with other highly cited review (Siemens & Baker, 2012). And finally, two good specific reviews about LA provide us with an introductory to this area (Ferguson, 2012) and an analysis of the citation networks of the area (Dawson, Gašević, Siemens, & Joksimovic, 2014).

3 | EDM/LA KNOWLEDGE DISCOVERY CYCLE

The process of applying EDM/LA is a cycle application of the general knowledge discovery and data mining (KDD) process (Figure 3) although there are some important differences with specific characteristics in each step as described in the subsections below.

3.1 | Educational environment

Depending on the type of the educational environment (traditional classroom education, computer-based or BL education) and the information system used, such as: LMS (Learning Management System), ITS (Intelligent Tutoring System), MOOC (Massive Open Online Course), etc., different kinds of data can be collected in order to resolve different educational problems (Romero & Ventura, 2013).

3.2 | Educational data

Educational data are gathered (Romero, Romero, & Ventura, 2014) from different sources such as the interaction between instructors, students and the educational (e.g., navigation behavior, input in quizzes, interactive exercises, forum messages, etc.) administrative data (e.g., school and teacher information), demographic data (e.g., gender, age, etc.), student affectivity (e.g., motivation, emotional states), and so on. Educational environments can store a huge amount of potential data from multiple sources with different formats and with different granularity levels (from coarse

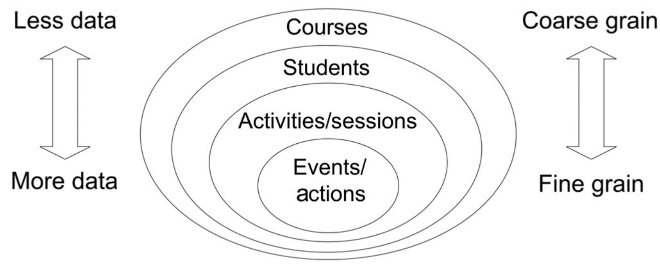


FIGURE 4 Different levels of granularity and their relationship to the amount of data

to fine grain) or multiple levels of meaningful hierarchy (keystroke level, answer level, session level, student level, classroom level, and school level) that provide more or less data (Figure 4). Gathering and integrating all this raw data for mining are nontrivial tasks on their own and thus a preprocessing step is necessary.

3.3 | Preprocessing

Data preprocessing is a hard and complicated task, and sometimes the data preprocessing itself takes up more than half of the total time spent on solving the data mining problem (Bienkowski et al., 2012). Educational data available (raw, original, or primary data) to solve a problem are not in the appropriate form (or abstraction). And so, it is necessary to convert the data to an appropriate form (modified data) for solving each specific educational problem. This includes choosing what data to collect, focusing on the questions to be answered, and making sure the data align with the questions. Traditional preprocessing tasks are applied to educational data with some specific issues (Romero et al., 2014) such as the next ones. Feature engineering for generating and selecting attributes/variables with information about the students is very important. Normally we can reduce and transform all available attributes into a summary table for better analysis. Continuous attributes are normally transformed/discretized into categorical attributes in order to improve their comprehensibility. Finally, it is important to maintain and protect the confidentiality of student by anonymizing data and deleting all personal information (not useful for mining purposes) such as name, e-mail, telephone number, and so on. In this line, we can have into consideration use guidelines about ethical issues, data privacy, informed consent, etc. when using educational data (Pardo & Siemens, 2014).

3.4 | Methods and techniques

The majority of traditional data mining techniques including but not limited to visualization, classification, clustering, and association analysis techniques have been already applied successfully in the educational domain (Baker, 2015). Nevertheless, educational systems have also some special characteristics (hierarchical and longitudinal data) that require a specific treatment of the mining problem and preprocess of the data. There are a wide range of EDM and LA methods and techniques used for solving different educational problems as described in the Methods section below.

3.5 | Interpretation and application of the new knowledge

Taking action is the ultimate goal of any learning analytics process and the results of follow-up actions will determine the success or failure of our analytical efforts (Siemens, 2013). So, the discovered new knowledge by the EDM/LA methods have to be used by instructors and academic authorities to make interventions and decisions in order to improve student learning performance. It is very important that the previous models obtained by the EDM/LA process were comprehensible in order to be useful for the decision-making process. In this line, white-box DM models such as decision trees are preferable to black-box models such as neural networks as they are more accurate but less comprehensible. Visualization techniques are also very useful for showing results in a way that is easier to interpret. Recommender systems are very useful for providing explanations and recommendations both to students and a nonexpert user in EDM/LA such as instructors.

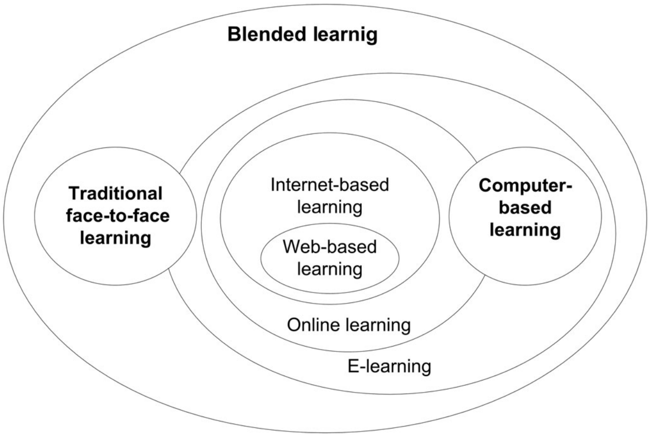


FIGURE 5 Types of educational environments and systems

TABLE 5 International levels of education

Level	Principal characteristics
Preprimary education	Designed for children from age 3 to the start of primary education.
Primary education or first stage of basic education	Normally starting between the ages of 5–7.
Lower secondary education or second stage of basic education	Designed for children from ages 8–14 to complete basic education.
Upper secondary education	More specialized education beginning at age 15 or 16 years.
Postsecondary nontertiary education	Programs that straddle the boundary between upper- and postsecondary education from an international point of view.
First stage of tertiary education	Tertiary programs having an educational content more advanced than those offered at previous levels.
Second stage of tertiary education	Tertiary programs leading to the award of an advanced research qualification, for example, Ph.D.

4 | EDUCATIONAL ENVIRONMENTS AND DATA

There is a wide variety of educational environments (Figure 5) such as traditional education, CBE, and BL. Each one of them provides different data sources (Romero & Ventura, 2007).

4.1 | Traditional face-to-face education

Traditional education or back-to-basics are the most widely used educational system, based mainly on face-to-face contact between educators and students organized through lectures, class discussion, small groups, individual seat work, and so on. Traditional education systems are classified in different levels by UNESCO[6] as we can see in Table 5. These systems gather information on student attendance, marks, curriculum goals, class, schedule information, and so on. Finally, it is important to note that all these traditional systems can also use computer-based educational systems as a complementary tool to face-to-face sessions.

4.2 | Computer-based educational systems

CBE means using computers in education to provide direction, to instruct or to manage instructions given to the student. CBE systems were originally simple stand-alone educational applications that ran on a local computer. But the

TABLE 6 Examples of computer-based educational systems

System	Description
Adaptive and Intelligent Hypermedia System (AIHS)	These systems attempt to be more adaptive by building a model of the goals, preferences, and knowledge of each individual student and using this model throughout interaction with the student in order to adapt to the needs of that student. The data recorded by AIHs are similar to ITS data.
Intelligent Tutoring System (ITS)	ITSs provide direct customized instruction or feedback to students by modeling student behavior and changing its mode of interaction with each student based on its individual model. Normally, it consists of a domain model, student model, and pedagogical model. ITSs record all student-tutor interaction (mouse clicks, typing, and speech).
Learning Management System (LMS)	Suites of software that provide course-delivery functions: Administration, documentation, tracking, and reporting of training programs, classroom and online events, e-learning programs, and training content. They record any student activities involved, such as reading, writing, taking tests, performing tasks in real, and commenting on events with peers.
Massive Open Online Course (MOOC)	It refers to a web-based class designed to support a large number of participants. It can deliver learning content online to any person who wants to take a course, with no limit on attendance. They store the same information that LMS.
Test and quiz system	The main goal of these systems is to measure the students' level of knowledge with respect to one or more concepts or subjects by using a series of questions/items and other prompts for the purpose of gathering information from respondents. They store a great deal of information about students' answers, calculated scores, and statistics.
Other types	Wearable learning systems, learning object repositories, concept maps, social networks, WIKIs, forums, educational and serious games, virtual and augmented reality systems, and so on. They store different type of information about the interaction with the students.

TABLE 7 Examples of specific Educational Data Mining/Learning Analytics tools

Name	URL	Description
DataShop	https://pslcdatashop.web.cmu.edu/	It provides both a central repository to secure and store research data, and a set of analysis and reporting tools.
GISMO	http://gismo.sourceforge.net/	Graphical interactive monitoring tool that provides useful visualization of students' activities in online courses to instructors.
Inspire	https://moodle.org/plugins/tool_inspire	Moodle Analytics API that provides descriptive and predictive analytics engine, implementing machine learning backends.
LOCO-Analyst	http://jelenajovanovic.net/LOCO-Analyst/	Tool aimed at providing teachers with feedback on the relevant aspects of the learning process taking place in a web-based learning environment.
Meerkat-ED	http://www.reirab.com/MeerkatED	Tool for analyzing students' activity in a course offered on computer-supported collaborative learning tools.
MDM Tool	http://www.uco.es/kdis/research/software/	Framework for apply some data mining techniques in Moodle 2.7 version.
Performance Plus	https://www.d2l.com/higher-education/products/performance/	Package for delivering powerful analytics tools to help administrators, educators, and learners save quality time while maximizing impact and driving success.
SNAPP	https://web.archive.org/web/20120321212021/http://research.uow.edu.au/learningnetworks/seeing/snapp/index.html	Tool that allows users to visualize the network of interactions resulting from discussion forum posts and replies.
Solutionpath StREAM	https://www.solutionpath.co.uk/	Real-time system that leverage predictive models to determine all facets of student engagement

Abbreviation: API, application programming interface.

TABLE 8 Educational Data Mining/Learning Analytics public datasets

Datasets	URL	Description
ASSISTments Competition Dataset	https://sites.google.com/view/assistmentsdatamining/home	Competition where data miners can try to predict an important longitudinal outcome using real-world educational data.
Canvas Network dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId = doi:10.7910/DVN/1XORAL	Deidentified data from Canvas Network open courses (running January 2014–September 2015), along with related documentation.
DataShop	https://pslcdatashop.web.cmu.edu/index.jsp?datasets=public	LearnSphere's DataShop provides a central repository to secure and store research ITS data and set of analysis and reporting tools.
Educational Process Mining Dataset	https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A + A + Learning+Analytics+Data+Set	Students' logs during sessions over a simulation environment in digital electronics.
HarvardX-MITx dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147	Deidentified data from the first year of MITx and HarvardX MOOC courses on the edX platform along with related documentation.
KDD Cup 2010 Dataset	https://pslcdatashop.web.cmu.edu/KDDCup/	Challenge to predict student performance on mathematical problems from logs of student interaction with ITS.
Learn Moodle dataset	https://research.moodle.net/158/	Anonymized data from the "Teaching with Moodle August 2016" course from learn.moodle.net.
MOOC-Ed Dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZZH3UB	Communications taking place between learners in two offerings of the Massively Open Online Course for Educators (MOOC-Eds).
NAEP Data Mining Competition 2019	https://sites.google.com/view/dataminingcompetition2019/dataset	Competition for measuring students' test taking activities, and helps develop and test evaluation methods for educational analysis.
NUS Multisensor Presentation Dataset	http://mmas.comp.nus.edu.sg/NUSMSP.html	It contains real-world presentations recorded in a multisensor environment.
Open University Learning Analytics Dataset	https://analyse.kmi.open.ac.uk/open_dataset	It contains data about courses, students and their interactions with Moodle for seven selected courses.
Student Performance Dataset	https://archive.ics.uci.edu/ml/datasets/Student+Performance	This data approach student achievement in secondary education of two Portuguese schools.
xAPI-Educational Mining Dataset	https://www.kaggle.com/aljarah/xAPI-Edu-Data	Students' Academic Performance Dataset collected from e-learning system called Kalboard 360.

global use of Internet and the application of artificial intelligence (AI) techniques have led to today's plethora of new web-based intelligent educational systems. Some examples of computer-based educational systems used currently are listed and described in Table 6.

4.3 | Blended learning systems

BL environments combine face-to-face instruction with computer-mediated instruction. The terms "blended learning," "hybrid learning," and "mixed-mode instruction" are often used interchangeably in research literature. Blended courses increased access, convenience, and it provides more flexibility and freedom compared to face to face courses by

TABLE 9 Most popular Educational Data Mining/Learning Analytics (EDM/LA) methods

Method	Goal/description	Key applications
Causal mining	To find causal relationship or to identify causal effect in data.	Finding what features of students' behavior cause learning, academic failure, drop out, and so on.
Clustering	To identify groups of similar observations.	Grouping similar materials or students based on their learning and interaction patterns.
Discovery with models	To employ a previously validated model of a phenomenon as a component in another analysis.	Identification of relationships among student behaviors and characteristics or contextual variables. Integration of psychometric modeling frameworks into machine-learning models.
Distillation of data for human judgment	To represent data in intelligible ways using summarization, visualization, and interactive interfaces.	Helping instructors to visualize and analyze the ongoing activities of the students and the use of information.
Knowledge tracing	To estimate student mastery of skills, employing both a cognitive model that maps a problem-solving item to the skills required, and logs of students' correct and incorrect answers as evidence of their knowledge on a particular skill.	Monitoring student knowledge over time.
Nonnegative matrix factorization	To define a matrix of positive numbers with student test outcome data that may be decomposed into a matrix of items and a matrix of student mastery of skills.	Assessment of student skills.
Outlier detection	To point out significantly different individuals.	Detection of students with difficulties or irregular learning processes.
Prediction	To infer a target variable from some combination of other variables. Classification, regression, and density estimation are types of prediction methods.	Predicting student performance and detecting student behaviors.
Process mining	To obtain knowledge of the process from event logs.	Reflecting students' behavior based on traces of their evolution through the educational system.
Recommendation	To predict the rating or preference a user would give to an item.	To make recommendations to students with respect to their activities or tasks, links to visits, problems or courses to be done, and so on.
Relationship mining	To study relationships among variables and to encode rules. Association rule mining, sequential pattern mining, correlation mining, and causal data mining are the main types.	Identifying relationships in learner behavior patterns and diagnosing student difficulties.
Statistics	To calculate descriptive and inferential statistics.	Analyzing, interpreting and drawing conclusions from educational data.
Social network analysis	To analyze the social relationships between entities in networked information.	Interpretation of the structure and relations in collaborative activities and interactions with communication tools.
Text mining	To extract high-quality information from text.	Analyzing the contents of forums, chats, web pages, and documents.
Visualization	To show a graphical representations of data.	To produce data visualizations that help communicate results of EDM/LA research to educators.
Non-negative matrix factorization	To define a matrix M of positive numbers with student test outcome data that may be decomposed into two matrices: Q, which represents a matrix of items, and S, which represents student mastery of skills.	Assessment of student skills.

TABLE 10 Example of users/stakeholders and objectives

User/stakeholders	Objectives
Learners or Students	Learners are interested in understanding student needs and methods to improve the learner's experience and performance.
Educators or Instructors	Educators attempt to understand the learning process and the methods they can use to improve their teaching methods.
Scientific Researchers	Researchers focus on the development and the evaluation of Educational Data Mining/Learning Analytics (EDM/LA) techniques for effectiveness.
Administrators or Academic Authorities	Administrators are responsible for allocating the resources for implementation in institutions.

transforming significant amount of face to face sessions online. These systems gather information about the two previous face-to-face and the computer-based systems.

5 | TOOLS AND DATASETS

Nowadays, there is a wide array of well-known general purpose tools and frameworks that can be used for the purposes of conducting EDM and LA research (Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2017) such as: Rapidminer, Weka, SPSS, Knime, Orange, Spark Lib, and so on. However, these tools aren't easy for educators to use due to they are required to select the specific method/algorithm to apply/use and to provide the appropriate parameters in advance in order to obtain good results/models. So, the educators must possess a certain amount of expertise in order to find the right settings (Romero & Ventura, 2013). A solution to this problem is the use of some of the specific EDM/LA available software tools (Table 7). However, they only work with specific data in order to solve specific educational problems.

Most of the EDM/LA researchers normally use their own data for solving their specific educational problems. But it is a hard and very time-consuming task to gather and preprocess educational data (Romero et al., 2014). So, another option is to use some of the public datasets that are currently available for free download in Internet as we show in Table 8.

We want to highlight DataShop (Koedinger et al., 2010) as one of the first and biggest dataset that also provides a tool for researching about ITS. As we see in Table 8, currently there are not many public datasets available and they are not from all the types of educational environments (most of them are from e-learning systems). So, we think that in the future it will be very useful to develop a specific EDM/LA datasets repository similar to the general UCI Machine Learning Repository.[7] It is important to note that these public datasets must be portable and must consider principles of data ethics, privacy, protection, and consent (Ferguson, Hoel, Scheffel, & Drachsler, 2016). The idea of data portability is that educational institution/instructors/researchers should not have their own data stored in "silos" or "walled gardens" that are incompatible with one another but to use standards such as Experience Application Programming Interface (xAPI)[8]or IMS Caliper.[9]

6 | METHODS AND APPLICATIONS

There is a wide range of popular methods (Table 9) within EMD and LA (Baker & Inventado, 2014; Bakhshinategh et al., 2018; Romero & Ventura, 2013) for solving educational problems or application. Most of these techniques are widely acknowledged to be universal across types of data mining, such as visualization, prediction, clustering, outlier detection, relationship mining, causal mining, social network analysis, process mining, and text mining. And others have more prominence within education, such as the distillation of data for human judgment, discovery with models, knowledge tracing, and non-negative matrix factorization.

However, the number of possible objectives or educational problems in EDM/LA is huge and this taxonomy does not cover all the possible tasks. In fact, there are many more specific objectives depending on the viewpoint of the final user. Although an initial consideration seems to involve only two main groups of potential users/stakeholders—the

TABLE 11 Some current applications or topics of interest of Educational Data Mining/Learning Analytics (EDM/LA) research community

Topics of interest	Description	Reference
Analyzing educational theories	To analyze how learning theories and learning analytics could be integrated in educational research.	(Wong et al., 2019)
Analyzing pedagogical strategies	To analyze and explore the application and effect of pedagogical strategies with EDM/LA techniques.	(Shen, Mostafavi, Barnes, & Chi, 2018)
Analyzing programming code	To apply EDM/LA techniques focused on analyzing code from programming courses, programming assignments/submissions, and so on.	(Li & Edwards, 2018)
Collaborative learning and teamwork group	To analyze collaborative learning and to predict the team grade in teamwork groups.	(Hernández-García, Acquila-Natale, Chaparro-Peláez, & Conde, 2018)
Curriculum mining/ analytics	To analyze program structure, course grading, and administrative curricular data in order to improve curriculum development, program quality, and so on.	(Hilliger, Miranda, Celis, & Pérez-SanAgustín, 2019)
DashBoards and visual learning analytics	To apply a visualization technique to explore and understand relevant user traces that are collected in (online) environments and to improve (human) learning.	(Millecamp, Broos, De Laet, & Verbert, 2019)
Deep learning	To apply neural network architectures with multiple layers of processing units in EDM/LA research area.	(Hernández-Blanco, Herrera-Flores, Tomás, & Navarro-Colorado, 2019)
Discovery causal relationships	To find causality relationship among attributes in an educational dataset.	(de Carvalho & Zarate, 2019)
Early warning systems	To predicting student's performance and students at risk as soon as possible in order to intervene early to facilitate student success.	(Cano & Leonard, 2019)
Emotional learning analytics	To study affect during learning and the importance of emotion to learning.	(D'Mello, 2017)
Evaluating the efficacy of interventions	To evaluate the efficacy of interventions, data-driven student feedback, actionable advice, and so on.	(Sonderlund, Hughes, & Smith, 2018)
Feature engineering methods	To build automatically attributes or students features using machine learning techniques.	(Botelho, Baker, & Heffernan, 2019)
Game learning analytics	To apply data-mining and visualization techniques to player interactions in serious games.	(Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, & Fernández-Manjón, 2019)
Interpretable and explanatory learner models	To develop "white box" interpretable, explanatory, usable, and highly comprehensible leaner models.	(Rosé, McLaughlin, Liu, & Koedinger, 2019)
Learning foreign language	To apply EDM/LA techniques for improving of foreign language learning.	(Bravo-Agapito, Frances, & Seaone, 2019)
Measuring self-regulated learning	To apply EDM/LA techniques to measure self-regulated learning feature and behaviors in students.	(ElSayed, Caeiro-Rodríguez, MikicFonte, & Llamas-Nistal, 2019)
Multimodal learning analytics	To apply of machine learning and increasingly affordable sensor technologies for providing new types learning insights that happen across multiple contexts.	(Spikol et al., 2017)

(Continues)

TABLE 11 (Continued)

Topics of interest	Description	Reference
Orchestrating learning analytics	To study the adoption, implications for practice, and other factors in ongoing LA adoption processes at classroom level.	(Prieto, Rodríguez-Triana, Martínez-Maldonado, Dimitriadis, & Gašević, 2019)
Providing personalized feedback	To generate personalized feedback automatically or semiautomatically to support the student learning.	(Pardo, Jovanovic, Dawson, Gašević, & Mirriahi, 2019)
Sentiment discovery	To automatically identify the underlying attitudes, sentiments, and subjectivity in learners and learning resources.	(Han, Wu, Huang, Huang, & Zhao, 2019)
Transfer learning	To develop models that can be transferable or applied to other similar courses/institutions/etc.	(Ding, Wang, Hemberg, & O'Reilly, 2019)
Understanding navigation paths	To discover process-related knowledge and navigational learning from event logs recorded by e-learning systems.	(Bogarín, Cerezo, & Romero, 2018)
Writing analytics	To apply text mining and analytics tools to text data from forums, chats, social networks, assessments, essays, and so on.	(Ferreira-Mello, André, Pinheiro, Costa, & Romero, 2019)

learners and the instructors—there are actually more groups involved with many more objectives, as can be seen in Table 10. And, in order to show more examples of the most promising applications of the EDM/LA, Table 11 shows some of the current hot topics or more interesting problems in the area.

7 | CONCLUSIONS AND FUTURE TRENDS

EDM and LA are two interdisciplinary communities of computer scientists, learning scientists, psychometricians, and researchers from other areas with the same objective of improve learning starting from data. This area has grown quickly in the last two decades with two different annual conferences (EDM and LAK), two specific journals (*JEDM* and *JLA*), and increasing number of books, papers, and surveys/reviews. Additionally, they are a current move from the lab to the general market for using EDM/LA by educational institutions and schools worldwide and it is expected that in 2020 all education research involves analytics and data mining (Baker and Inventado, 2014). All this indicates us that EDM/LA will become soon a mature area that will be widely used not only by researchers but also by instructors, educational administrators, and related business from all over the world. EDM/LA has impacted our understanding of learning and produced insights that have been translated to mainstream practice or contributed to theory. The research in this area has developed in the study focus and sophistication of analyses, but the impact on practice, theory, and frameworks have been more limited (Dawson, Joksimovic, Poquet, & Siemens, 2019). So, it is necessary the stimulus from existing research organizations (such as SoLAR and IEDM), funding agencies, collaborations, and the active promotion of established works such as the LACE initiative[10] in order to increase the impact on practice and to move from exploratory models to more holistic and integrative systems-level research.

As for future trends in the area, our previous survey (Romero & Ventura, 2013) identified two trends. However, one of them has not been completely achieved yet and the other remains a challenge. The first one was that it is necessary more freely available EDM tools in order to a wider and broader population can use them. As we can see in section Tools and Datasets, currently there is now a wide array of specific purpose tools. However, one still needs to develop general purpose EDM/LA tools to apply several tasks for solving different educational problems from the same interface/tool. It is also necessary to improve the portability of the obtained models from these tools. The second one was that educators and institutions should develop a data-driven culture of using data for making instructional decisions and improving instruction. However, according to a recent report (Joksimović, Kovanović, & Dawson, 2019), the majority of the institutions continue aware of the benefits provided by the analysis of large-scale data about student learning. In order to address the complexity of scaling learning analytics Dawson et al. (2018) argued for the inclusion of new forms of leadership models in education to stimulate and promulgate systemic change. And specific Learning Analytics challenges to overcome in higher education institutions are proposed and grouped into the next seven categories (Leitner, Ebner, & Ebner, 2019):

1. Purpose and gain. It is necessary to make the goals of the LA initiative transparent, clarifying exactly what is going to happen with the information and explicitly what is not.
2. Representation and actions. To choose the right environment for the learner's feedback, the correct visualization technique to provide recommendations and results to the students.
3. Data. A policy needs to be created for LA that aligns with the organization's core principles. Transparent communication about where the data are stored, what is being done to ensure data security and privacy and how the data are evaluated and used (and by whom) is essential.
4. IT infrastructure. Efforts should be made from the beginning to search for possible solutions to set up the necessary internal or external IT infrastructure and contact and establish connections with relevant people.
5. Development and operation. Scalability is maybe one of the most frequently underestimated problems in today's IT industry. A distinction must be made as to whether processes have to be carried out manually, semiautomatically or fully automatically.
6. Privacy. All LA implementations have to ensure the privacy of the involved parties. The general lifetime of personal data is a topic that requires further discussion.
7. Ethics. LA implementers must find a suitable way to meet high ethical standards and ensure a beneficial outcome for all stakeholders.

Additionally, the Baker Learning Analytics Prizes (BLAP) proposes the next six specific research problems as challenges of EDM/LA area (Baker, 2019):

1. Transferability: The (learning system) Wall Transfer student model from learning system A to learning system B. Improve an already-good student model in learning system B. Change behavior of learning system B in runnable fashion.
2. Effectiveness: Differentiating Interventions and Changing Lives Publicize criterion for intervention. Assign students to control or experimental group. Use analytics, only within experimental group, to assign intervention collect longer-term outcome measure. Demonstrate that experimental/analytics-intervention group performs better than experimental/analytics-no-intervention group. But that experimental/analytics-no-intervention group does not perform better than control/analytics-no-intervention group.
3. Interpretability: Instructors Speak Spanish, Algorithms Speak Swahili. Take a complex model of a learning analytics phenomenon. Develop a no-human-in-the-loop method of explaining the model. Present the explanation to five (new) data scientists and users. Ask the participants to explain what decision the model will make, and why, for five case studies. Code the explanations of the model's decisions. Verify if the data scientists and users interpret the model the same way for the case studies.
4. Applicability: Knowledge Tracing Beyond the Screen. Take data from at least four students completing learning activity together. Model at least four distinct skills for each student. Predict immediate future performance for these skills.
5. Generalizability: The General-Purpose Boredom Detector. Build an automated detector of affect. Demonstrate that the detector works for an entirely new learning system with different interactions and with AUC ROC $\geq$0.65.
6. Generalizability: The New York City and Marfa Problem. Build an automated detector for a commonly-seen outcome or measure. Collect a new population distinct from the original population. Demonstrate that the detector works for the new population with degradation of quality under 0.1 (AUC ROC, Pearson/Spearman correlation) and remaining better than chance.

Finally, we want to propose some personal visionary ideas that, in our opinion, might form very promising trends of EDM/LA in a near future:

- Taking into account all students' personal data through their whole life. Currently, information considered in EDM/LA is mainly based only in the interaction of students with a single educational environment, but in a near future thanks to big data and internet of things (IoT) (Al-Emran, Malik, & Al-Kabi, 2020), we will be able to have information available for each student from their birth to this very moment and on real time. It will imply the integration of not only the traditional performance and usage data gathered from all the previous institutions and educational environments each student has used, but also the information about the personal status of each student from different points of view such as medical, familiar, economical, religious, sexual, relationship, emotional, psychological, and so on. All these data could be gathered from multiple available sources and they could be fusioned (Ding, Jing, Yan, & Yang, 2019) in order to be used for improving and personalizing the learning process to each single student in each specific moment of their live to a new level of precision.

- Applying and integrating EDM/LA to upcoming technological educational environments. Over the last decade the great advances on innovative technologies has enabled the development of new educational systems from mobile and ubiquitous to virtual reality, augmented reality environments, and holograms (Cerezo, Calderón, & Romero, 2019). In the next few decades, quantum leaps will be associated to the application of AI. In this context, it is not wrong to think that instructors could be replaced by machines without students noticing the change thanks to current progresses in intelligent humanoids robots (Newton & Newton, 2019) and conversational agents or voice assistant interfaces in educational environments (Kloos, Catálan, Muñoz-Merino, & Alario-Hoyos, 2018). But these systems will need EDM/LA techniques for analyzing tons of data and generating portable analytics models[11] in real time in order to address the specific forthcoming educational challenges of these future technological environments.
- Analyzing and mining data directly gathered from students' brain for a better understanding of the learning. The brain is the key factor to really understand how students learn. The promising advances in human neuroscience and pervasive neurotechnology (brain-computer interfaces, BCIs) are giving rise to unprecedented opportunities for getting, collecting, sharing, and manipulating any kind of information gathered from the human brain (Williamson, 2019). In a near future, these intimate data about student's psychological state and neural activity could be analyzed and mining in real-time thanks to upcoming small high-quality electroencephalography (EEG) devices. These brain data, together with other multimodal data (Giannakos, Sharma, Pappas, Kostakos, & Velloso, 2019) could be integrated and used by EDM/LA researchers in order to achieve a fully understanding of students' learning process.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Cristobal Romero Morales and **Sebastian Ventura**: Investigation; writing-original draft, review, and editing.

ORCID

Cristobal Romero https://orcid.org/0000-0003-4180-4948
Sebastian Ventura https://orcid.org/0000-0003-4216-6378

ENDNOTES

[1] http://www.columbia.edu/~rsb2162/bigdataeducation.html.

[2] http://www.educationaldatamining.org/JEDM/.

[3] https://solaresearch.org/stay-informed/journal/.

[4] https://online-journals.org/index.php/i-jai.

[5] https://solaresearch.org/stay-informed/journal/.

[6] http://uis.unesco.org/en/topic/international-standard-classification-education-isced.

[7] https://archive.ics.uci.edu/ml/index.php.

[8] https://www.adlnet.gov/projects/xapi/.

[9] https://www.imsglobal.org/activity/caliper.

[10] http://www.laceproject.eu/lace/.

[11] http://dmg.org/.

RELATED WIREs ARTICLES

Data mining in education
Educational data science in massive open online courses
A survey on educational process mining
Data mining in foreign language learning
Text mining in education

FURTHER READING

Mostow, J., & Beck, J. (2006). Some useful tactics to modify, map and mine data from intelligent tutors. *Natural Language Engineering*, *12*(2), 195–208.

REFERENCES

Al-Emran, M., Malik, S. I., & Al-Kabi, M. N. (2020). A survey of internet of things (IoT) in education: Opportunities and challenges. In *Toward social internet of things (SIoT): Enabling technologies, architectures and applications* (pp. 197–209). Cham, Switzerland: Springer.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 1–14.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International conference on learning analytics and knowledge*, Vancouver, Canada (pp. 267–270).

Baker, R. S. (2015). *Big data and education* (2nd ed.). New York, NY: Teachers College, Columbia University.

Baker, R. S. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. *Journal of Educational Data Mining*, *11*(1), 1–17.

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*(1), 3–17.

Baker, R. S. J.d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice*. Berlin, Germany: Springer.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537–553.

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *Computer Science*, *2*(6), 63–69.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief* (pp. 1–57). Washington, DC: U.S. Department of Education, Office of Educational Technology.

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *WIREs: Data Mining and Knowledge Discovery*, *8*(1), e1230.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2019). Machine-learned or expert-engineered features? Exploring feature engineering methods in detectors of student behavior and affect. In *The twelfth international conference on educational data mining*, Montréal, Canada.

Bravo-Agapito, J., Frances, C., & Seaone, I. (2019). Data mining in foreign language learning. *WIREs: Data Mining and Knowledge Discovery*, *10*(1), e1287.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *Educause Review*, *42*(4), 40.

Cano, A., & Leonard, J. (2019). Interpretable multi-view early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, *12*, 198–211.

Cerezo, R., Calderón, V., & Romero, C. (2019). A holographic mobile-based application for practicing pronunciation of basic English vocabulary for Spanish speaking children. *International Journal of Human-Computer Studies*, *124*, 13–25.

Custer, S., King, E. M., Atinc, T. M., Read, L., & Sethi, T. (2018). *Toward data-driven education systems: Insights into using information to measure results and manage change*. Washington, DC: Center for Universal Education at the Brookings Institution.

Daniel, B. K. (2019). Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, *50*(1), 101–113.

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28.

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics area. In *Proceedings of the fourth international conference on learning analytics and knowledge*, Indiana, USA (pp. 231–240).

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. In *Proceedings of the 9th international conference on learning analytics & knowledge*, Tempe, Arizona (pp. 446–455).

Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., & Gasevic, D. (2018). Rethinking learning analytics adoption through complexity leadership theory. In *Proceedings of the 8th international conference on learning analytics and knowledge*, Sydney, Australia (pp. 236–244).

de Carvalho, W. F., & Zarate, L. E. (2019). Causality relationship among attributes applied in an educational data set. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 1271–1277). Limassol, Cyprus: ACM.

Ding, M., Wang, Y., Hemberg, E., & O'Reilly, U. M. (2019). Transfer learning using representation learning in massive open online courses. In *Proceedings of the 9th international conference on learning analytics & knowledge*, Tempe, Arizona (pp. 145–154).

Ding, W., Jing, X., Yan, Z., & Yang, L. T. (2019). A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion. *Information Fusion*, *51*, 129–144.

D'Mello, S. (2017). Emotional learning analytics. In *Handbook of learning analytics* (p. 115). New York, NY: SOLAR.

Dormezil, S., Khoshgoftaar, T., & Robinson-Bryant, F. (2019). *Differentiating between educational data mining and learning analytics: A bibliometric approach*. LABBEC Workshop (Learning analytics: Building bridges between the Education and the Computing communities; pp. 1–6), Montreal, Canada.

ElSayed, A. A., Caeiro-Rodríguez, M., MikicFonte, F. A., & Llamas-Nistal, M. (2019). Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review. In *World conference on mobile and contextual learning*, Delft, Netherlands (pp. 46–53).

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4* (5/6), 304–317.

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest editorial: Ethics and privacy in learning analytics. *SoLAR*, *3*(1), 5–15.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs: Data Mining and Knowledge Discovery*, *9*(6), e1332.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*, 108–119.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, *15*(42), 42–57.

Han, Z., Wu, J., Huang, C., Huang, Q., & Zhao, M. (2019). A review on sentiment discovery and analysis of educational big-data. *WIREs: Data Mining and Knowledge Discovery*, *10*(1), e1328.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*, 1–22.

Hernández-García, Á., Acquila-Natale, E., Chaparro-Peláez, J., & Conde, M. Á. (2018). Predicting teamwork group assessment using log data-based learning analytics. *Computers in Human Behavior*, *89*, 373–384.

Hilliger, I., Miranda, C., Celis, S., & Pérez-SanAgustín, M. (2019). Evaluating usage of an analytics tool to support continuous curriculum improvement. In *EC-TEL practitioner proceedings*, Delf, Netherland (pp. 1–14). Delft, Netherlands.

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics. *HERDSA Review of Higher Education*, *6*, 27–63.

Kloos, C. D., Catálan, C., Muñoz-Merino, P. J., & Alario-Hoyos, C. (2018). Design of a conversational agent as an educational tool. In *Learning with MOOCS (LWMOOCS)*, Madrid, Spain (pp. 27–30).

Koedinger, K. R., Baker, R. S. J. D., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC datashop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research. New York, NY: SOLAR.

Leitner, P., Ebner, M., & Ebner, M. (2019). Learning analytics challenges to overcome in higher education institutions. In *Utilizing learning analytics to support study success* (pp. 91–104). Cham, Switzerland: Springer.

Li, Z., & Edwards, S. (2018). Applying recent-performance factors analysis to explore student effort invested in programming assignments. In *Proceedings of the international conference on frontiers in education: Computer science and computer engineering (FECS)* (pp. 3–10).

Liñán, L. C., & Pérez, Á. A. J. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, *12*(3), 98–112.

Millecamp, M., Broos, T., De Laet, T., & Verbert, K. (2019). DIY: Learning analytics dashboards. In *Companion proceeding of the 9th international conference on learning analytics & knowledge (LAK'19)*, Tempe, Arizona. (pp. 947–954).

Newton, P., & Newton, L. (2019). When robots teach: Towards a code of practice. *Research in Education*, 1–7.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Prieto, L. P., Rodríguez-Triana, M. J., Martínez-Maldonado, R., Dimitriadis, Y., & Gašević, D. (2019). Orchestrating learning analytics (OrLA): Supporting inter-stakeholder communication about adoption of learning analytics at the classroom level. *Australasian Journal of Educational Technology*, *35*(4), 14–33.

Prieto, L. P., Sharma, K., Dillenbourg, P., & Jesús, M. (2016). Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors. In *Proceedings of the sixth international conference on learning analytics & knowledge*, Edinburgh, United Kingdom (pp. 148–157).

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In *Educational data mining* (pp. 29–64). Cham, Switzerland: Springer.

Romero, C., & Ventura, S. (2006). *Data mining in E-learning*. Southampton, England: Wit-Press.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Journal of Expert Systems with Applications*, *1*(33), 135–146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *40*(6), 601–618.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *WIREs: Data Mining and Knowledge Discovery*, *7*(1), e1187.

Romero, C., Ventura, S., Pechenizky, M., & Baker, R. (2010). *Handbook of educational data mining*. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Editorial Chapman and Hall/CRC Press, Taylor & Francis Group.

Romero, C., Ventura, S., & Salcines, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384.

Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, *50*(6), 2943–2958.

Shen, S., Mostafavi, B., Barnes, T., & Chi, M. (2018). Exploring induced pedagogical strategies through a Markov decision process framework: Lessons learned. JEDM|. *Journal of Educational Data Mining*, *10*(3), 27–68.

Siemens, G., Baker, R.S.J.d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, Vancouver, Canada (pp. 1–3).

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, *46*(5), 30.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106.

Sonderlund, A., Hughes, E., & Smith, J. R. (2018). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618.

Spikol, D., Prieto, L. P., Rodríguez-Triana, M. J., Worsley, M., Ochoa, X., Cukurova, M., & Ringtved, U. L. (2017). Current and future multimodal learning analytics data challenges. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 518–519). Vancouver, Canada: ACM.

Williamson, B. (2019). Brain data: Scanning, scraping and sculpting the plastic learning brain through neurotechnology. *Postdigital Science and Education*, *1*(1), 65–86.

Wong, J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., & Paas, F. (2019). Educational theories and learning analytics: From data to knowledge. In *Utilizing learning analytics to support study success* (pp. 3–25). Cham, Switzerland: Springer.

How to cite this article: Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining Knowl Discov*. 2020;10:e1355. https://doi.org/10.1002/widm.1355