Education and Information Technologies (2023) 28:8299–8333
https://doi.org/10.1007/s10639-022-11536-0

Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)

Nabila Sghir[1] · Amina Adadi[1] · Mohammed Lahmer[1]

Received: 13 September 2022 / Accepted: 13 December 2022 / Published online: 20 December 2022

Abstract

The last few years have witnessed an upsurge in the number of studies using Machine and Deep learning models to predict vital academic outcomes based on different kinds and sources of student-related data, with the goal of improving the learning process from all perspectives. This has led to the emergence of predictive modelling as a core practice in Learning Analytics and Educational Data Mining. The aim of this study is to review the most recent research body related to Predictive Analytics in Higher Education. Articles published during the last decade between 2012 and 2022 were systematically reviewed following PRISMA guidelines. We identified the outcomes frequently predicted in the literature as well as the learning features employed in the prediction and investigated their relationship. We also deeply analyzed the process of predictive modelling, including data collection sources and types, data preprocessing methods, Machine Learning models and their categorization, and key performance metrics. Lastly, we discussed the relevant gaps in the current literature and the future research directions in this area. This study is expected to serve as a comprehensive and up-to-date reference for interested researchers intended to quickly grasp the current progress in the Predictive Learning Analytics field. The review results can also inform educational stakeholders and decision-makers about future prospects and potential opportunities.

Keywords Learning analytics · Educational data mining · Predictive modelling · Machine learning · Higher education

✉ Nabila Sghir
nabila.sghir@edu.umi.ac.ma; nabila.sghir@gmail.com

Amina Adadi
a.adadi@umi.ac.ma

Mohammed Lahmer
m.lahmer@umi.ac.ma

[1] Moulay Ismail University, Meknes, Morocco

1 Introduction

Over the last years, the growth in the use of Artificial Intelligence (AI) accelerated the development of the capacity of capturing and gaining insight into data on various aspects of the learning experience. This has resulted in the emergence of fields that study learners' data, such as Learning Analytics (Nunn et al., 2016) and Educational Data Mining (Dutt et al., 2017). Recent developments in these fields have attracted much attention from researchers and practitioners, as well as diverse stakeholders, who are interested in exploring these data-driven technologies to enhance learning, increase and improve students' performance, and address potential issues in higher education such as students' retention and dropout. One of the prominent techniques to achieve these goals is predictive modelling (Brooks & Thompson, 2017). With predictive analytics, it is possible to create forecasts for the future by analyzing past trends in learning experiences. During the last ten years, efficient and sophisticated predictive models developed with machine and deep learning have allowed us to discover complex hidden characteristics in data. It also permitted us to substantially push forward the achievable prediction accuracy. Predictive models have thus gained popularity in education as a competitive strategy that goes beyond simple monitoring of student performance, it allows anticipatory management of student success and early design of preventive intervention measures for students at risk. This popularity is illustrated by a growing body of research addressing learning issues by using predictive models.

This study aims to perform a comprehensive systematic review of this body of research to evaluate the current progress, trends, arising challenges, and future research avenues related to Predictive Learning Analytics (PLA). Specifically, we intend to identify, discuss, compare, and contrast the most recent and relevant research papers on the topic of PLA in higher education in order to determine the relevance of this technique and its effect on the learning process, as well as how predicting students' outcomes can positively contribute to students' success. Furthermore, we want to know what kinds of features are relevant for predicting student outcomes, as well as what features work for what outcomes. From a technical standpoint, we want to define the characteristics to consider in order to build high-performing predictive models, identify the type and size of data to use, choose the algorithm or group of algorithms to employ, and finally specify the metrics used to evaluate the predicting models. These findings would be useful for future researchers or practitioners that intend to implement Learning Analytics systems that use prediction to improve teaching and learning.

The study was developed under the systematic review approach and considered the period 2012—2022 to highlight the progression from traditional predictive models toward modern machine and deep learning models. This period will also allow us to investigate the mutual impact of the COVID-19 pandemic and PLA.

This paper is structured as follows: in Sect. 2 we provide a narrative description of the main concepts addressed in this study and a comparative analysis of related reviews. Section 3 describes the methodology used in this review and gives a detailed overview of the process followed to select relevant papers. The findings and the results of this review are presented in Sect. 4. Section 5 provides a discussion of the challenges and future research directions; it also highlights the limitations of this study. Finally, the "Conclusion" section concludes this review.

2 Background

This section presents a narrative description of the main concepts addressed in this review, including (i) Learning analytics, (ii) Educational Data Mining, and (iii) Predictive Modelling. We also present an analysis of the main related works.

2.1 Artificial intelligence in education

AI is set to revolutionize key industries and vital sectors. Indeed, rapidly increasing computing power and connectedness have made it possible to compile, analyze and share large volumes of valuable data, which is now more accessible than ever before. In education, although the application of AI in education (AIEd) has been the subject of research for about 30 years (Zawacki-Richter et al., 2019). On a practical scale, AI research in the field is still considered in its infancy. It is only in the last decade that educators have started to implement AI-based solutions to support students' learning experiences and teachers' practices (Zawacki-Richter et al., 2019), with a notable digital divide between developed and developing countries (Pedró et al., 2019). AIED is best represented by two data-driven fields that have emerged in the last few years, namely (i) Learning Analytics and (ii) Educational Data Mining.

Learning Analytics (LA) LA is generally 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" (Siemens & Long, 2011), in a nutshell, LA is a tool to enhance learning by analyzing data. Based on logs and trace data collected from student and teacher activities, LA can produce support for: (i) learners, by receiving more meaningful and timely feedback or even self-monitoring their progress through LA data, (ii) educators by allowing them to evaluate the efficacy of their instructions and make necessary enhancement to meet the needs of their students, and (iii) decision-makers by offering them useful suggestions to enhance their productivity and competitivity. LA is a multi-disciplinary field that involves techniques of data collection, analysis, visualization, and interpretation (El Alfy et al., 2019). As a data-driven field, the emergence and success of LA in recent years can be attributed to the rise of digital learning environments which increased the quality and access to educational data. Other technological innovations have also enabled LA, including the rapid development of the Internet, mobile technologies, and the democratization of the use of big data tools.

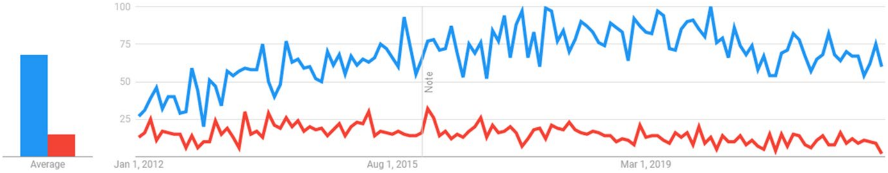

Fig. 1 "Learning analytics" (in blue) and "Educational Data Mining" (in red) search trends, 2012–2022. Data source: Google Trends (https://www.google.com/trends)

Educational Data Mining (EDM) EDM can be considered as the sister research field of LA. The International Educational Data Mining Society defines EDM as (Dutt et al., 2017) "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in". EDM is mainly concerned with applying machine learning, statistics, and data mining algorithms to educational attributes (Dutt et al., 2017). Its goal is to process raw educational data and extract valuable information to automate learning processes or intervention implementations (Du et al., 2020).

EDM and LA share similar goals and yet each draws its own unique research spectrum. In the literature, many scholars call for the need of defining a clear boundary between the two fields of LA and EDM (Rienties et al., 2020). Generally, EDM is identified as a technical-focused field whereas LA focuses on learning and educational issues (Baek & Doleck, 2021). Arguably, because it puts forth pedagogy/education matters, LA is attracting more interest compared to EDM. Figure 1 illustrates the recent interest in the two fields using google trends.

In this study, we do not intend to shed light on differences between LA and EDM, rather we aim to study the intersections and overlaps between the two fields. We admit that they have different focuses, but we believe that a synergetic approach will lead us to address the issue from both technical and educational perspectives. In this sense, we are interested in investigating how EDM methods, more precisely relationship-mining and prediction methods, are used in LA to understand learning dimensions including pedagogical, social, psychological, and organizational dimensions.

2.2 Predictive modelling in education

In both EDM and LA, predictive modelling has become a core research practice. It provides foresight and vision for the future that can help enhance learning effectiveness and prompt remedial, timely, and appropriate actions. Practically, predictions are produced by analyzing historical data and projecting it on a model generated to forecast likely outcomes with a certain accuracy. To create a predictive model, Brooks and Thompson (2017) described the main steps to follow, beginning with (i) identifying the problem, (ii) collecting the data required for analysis, and defining the predicted outcome, (iii) selecting the predictor variables that perfectly correlate with the chosen output, and finally (vi) building the predictive model using one or more algorithms.

Several algorithms are used to develop predictive models (Kumar et al., 2018), the most popular are: Decision Tree, Bayesian Classifier, Neural Networks, Support Vector Machine, K-Nearest Neighbor, and Logistic and Linear Regression. Algorithms are chosen based on the problem type, the nature of the outcome to be predicted, and the variables employed in the prediction. A common practice used by researchers is to test multiple algorithms and compare their performance to determine which one provides the most accurate prediction.

2.3 Related reviews

Data analytics in the educational field for the purpose of making predictions or other tasks holds a significant record of active research. Indeed, with the emergence of EDM and LA, several reviews were conducted in order to provide a broader overview of the results of analyzing data to deal with educational issues. In this subsection, we present an overview of the related reviews proposed in the literature and emphasize the significant differences between this work and the existing reviews.

The existing reviews differed largely in terms of the scope and the perception of the subject. For instance, Nunn et al. (2016) described the methods, benefits, and challenges of LA to apply it more effectively to improve teaching and learning in higher education. Other reviews focused on both EDM and LA. For example, Baek and Doleck (2021) showed the similarities and differences in research across the two fields by examining data analysis tools, common keywords, theories, and definitions.

Aside from reviews that discussed LA and EDM generally, some reviews focused on specific learning issue. An example is mentioned by de Oliveira et al. (2021), the authors focused on retention and dropout of higher education students and how LA can be used to help prevent these cases. Meanwhile, Namoun and Alshanqiti (2021) explored the prediction of student academic performance and presented the intelligent techniques used in performance prediction. Since deploying the right interventions to help students at risk of underperformance or discontinuation is important, some reviews focused on the effectiveness of interventions based on predictive models to help institutions implement the right interventions (Larrabee Sønderlund et al., 2019).

The identified literature reviews suffer from some notable limitations, some works examined LA and EDM in a broader context, whilst others had a tight focus on specific learning issue. Other reviews discussed the field from an educational point of view without addressing the technical aspect. Table 1 provides an overview of the limitations of the existing reviews related to LA, EDM, and predictive modelling.

Since the advent of deep learning methods and the outbreak of the COVID-19 pandemic, PLA has attained new interest and several advances have been made in the field that justify a new and updated review. Our aim is to overcome the limitations spotted in the existing body of reviews; hence, the primary contribution of this review is to organize and categorize the growing literature on PLA holistically and comprehensively, without focusing on a specific educational issue. In addition, we

Table 1 Related literature review description

Review	Focus	Limitations
El Alfy et al. (2019)	addressed the benefits that LA can provide, and the challenges regarding the use of it	- Offers a theoretical overview with no mention of methodologies - Missing paper filtration criteria - No focus on predictive modelling
Baek and Doleck (2021)	compared the similarities and differences in research across the two fields LA and EDM	- Only one database was used in the collection of studies (WoS) - Literature between 2015 and 2019 (not updated) - Does not focus on Machine Learning (ML) algorithms
de Oliveira et al. (2021)	Analyzed how LA can be used to help prevent failure case	- Limited to dropout management
Namoun and Alshanqiti (2021)	presented a fundamental understanding of the intelligent techniques used for the prediction of student performance	- Limited to predicting performance
Larrabee Sønderlund et al. (2019)	Reviewed the evidence in terms of the efficacy of LA-based interventions targeting academic retention, underperformance, and dropout rates	- Limited to LA interventions - Literature up to 2018(not updated)
Gasevic et al. (2019)	proposed an approach that can be used in the adoption of LA in higher education	- Systematic review guidelines were not followed - No focus on methods used in the prediction
Umer et al. (2021)	presented a review of prior studies that have utilized ML techniques to predict student performances by using historical data	- Focuses only on studies that predict student performance and dropout - Limited to blended learning
Liz-Domínguez et al. (2019)	provided an overview of the current state of research activity regarding predictive analytics in higher education, highlighting the Early Warning Systems (EWS) that have been used in practice	- Focuses on EWS - Literature up to March 2019 (not updated)
Chan et al. (2019)	examined the use of e-LA data in health care studies with regards to how the analytics is reported	- Limited to healthcare-related educational disciplines - No focus on predictive modelling
Rastrollo-Guerrero et al. (2020)	analyzed the modern techniques widely applied for predicting students' performance	- Limited to predicting students' performance

included recent research works to explore the brand-new methods and algorithms that emerged in predictive modelling. In that way, we hope to cover the subject from both a technical and educational perspectives. Investigating whether and how PLA was employed to assist the educational ecosystem during the pandemic's outbreak is also a contribution of this work. The detailed research methodology followed in this review as well as the findings are explained in the following sections.

3 Method

3.1 Search process and research questions

This systematic review was conducted with the aim of obtaining sufficient data to identify clusters of research in the literature related to the use of PLA to support students' learning process in higher education. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009) to analyze, evaluate and interpret relevant studies to answer our research questions.

In this review, we considered these main steps: (i) frame the research questions, (ii) define search terms, (iii) identify relevant studies based on inclusion and exclusion criteria, (iv) review publications in three rounds (title and keywords, abstract, and full-text screening), (v) extract relevant information, (vi) synthesize and interpret the findings. In order to guide this systematic review, the following research questions were formulated:

RQ1: What are the research trends of the PLA field in the last years?
RQ2: How does PLA contribute to supporting students' learning process?
RQ3: What are the ML methods used in conducting PLA?

By exploring these three research questions, we would like also to verify two main facts:

(1) PLA has been considered and applied effectively during the pandemic.
(2) Earlier related works (Baker & Inventado, 2014; Coelho & Silveira, 2017) claimed that deep neural networks (DNN) are not a typical method of choice in EDM and LA. Arguably the two communities prefer the use of traditional ML. To what extent this fact is still true in light of the recent advancement in neural network models?

3.2 Search strategy

A query-based search was carried out to select relevant papers in accordance with our research questions. To acquire the best possible results, a granular search query was formed by the union of (i) the "Predictive Learning analytics" keyword and a set of related terms, and (ii) a variation of keywords associated with "Higher education"

Table 2 Search query keywords

PLA related keywords	Higher education related keywords
Learning analytics, Predictive analytics, Predictive modelling, Predict, Forecast, Educational Data Mining, EDM, Artificial Intelligence in Education, AIED, Artificial Intelligence, Machine Learning, Data Mining	Higher Education, University, College, Learning outcomes, Learning results, Student outcomes, Student results, Student success, at risk, drop-out, grade, performance, enrollment, retention, fail, satisfaction, motivation, engagement

environment. Table 2 describes the keywords used to build the search query. The identified search terms were then compiled into a unified query using the "OR" and "AND" operators, to link, respectively, terms variation of the same group and terms of the two keyword clusters namely PLA and Higher Education.

To conduct our electronic searches, two main academic databases were selected: Scopus and Web of Science. Both are generally recognized as the most reliable scientific databases with a large use worldwide, in addition, most publication avenues led by the Society of Learning Analytics Research – SoLAR,[1] the International Educational Data Mining Society – IEDMS[2], and the International Artificial Intelligence in Education Society –IAIED,[3] three renowned references in respectively LA, EDM, and AIEd research, are available in these two databases. Google Scholar was also used as a secondary source to find additional research resources.

The search was performed on March 1st, 2022. Since our goal is to review the recent PLA literature related to the use of machine and deep learning specifically, we limited our search to the last decade (from 2012 to 2022). This time window will allow us to analyze both the impact of the upsurge of DNN and the impact of the COVID-19 pandemic.

3.3 Inclusion Criteria

The following inclusion criteria were applied to select relevant studies:

(i) The paper is written in English
(ii) Full text is available
(iii) The paper is peer-reviewed
(iv) Review, meta-analysis, survey, or commentary articles were excluded from the results
(v) The study is conducted in a higher education context
(vi) At least one ML method is used
(vii) The learning outcomes to be predicted are clearly stated
(viii) Data collection and preparation are clearly described

[1] https://www.solaresearch.org

[2] https://educationaldatamining.org/

[3] https://iaied.org/about/

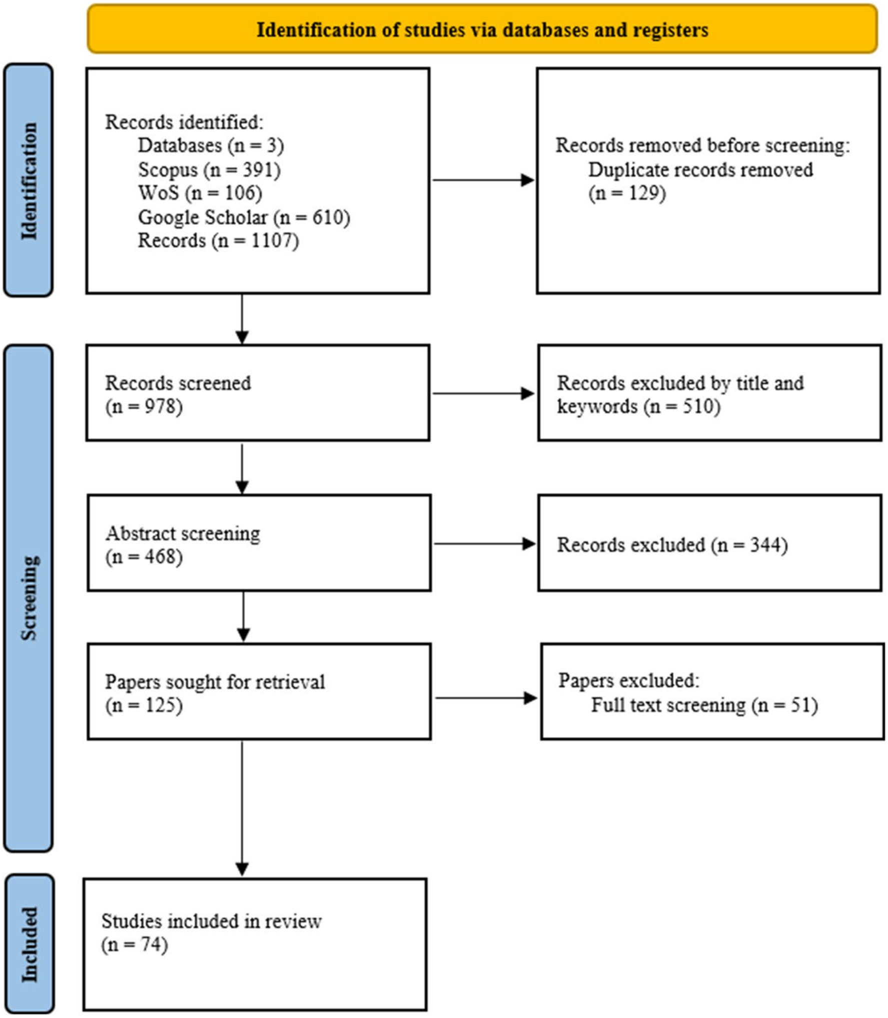
Identification of studies via databases and registers
Identification
Screening
Included
Records identified:
Databases (n = 3)
Scopus (n = 391)
WoS (n = 106)
Google Scholar (n = 610)
Records (n = 1107)
Records removed before screening:
Duplicate records removed
(n = 129)
Records screened
(n = 978)
Records excluded by title and keywords (n = 510)
Abstract screening
(n = 468)
Records excluded (n = 344)
Papers sought for retrieval
(n = 125)
Papers excluded:
Full text screening (n = 51)
Studies included in review
(n = 74)

Fig. 2 Search and selection of papers flow chart

(ix) The research results are properly validated, and the evaluation measures are clearly defined

Data from 1107 publications were retrieved and imported to Zotero,[4] a free reference management tool to collect and organize research. The search results were examined in three rounds against the inclusion criteria. Firstly, after screening the title and keywords, the publications that were clearly unrelated to the topic were

[4] https://www.zotero.org/

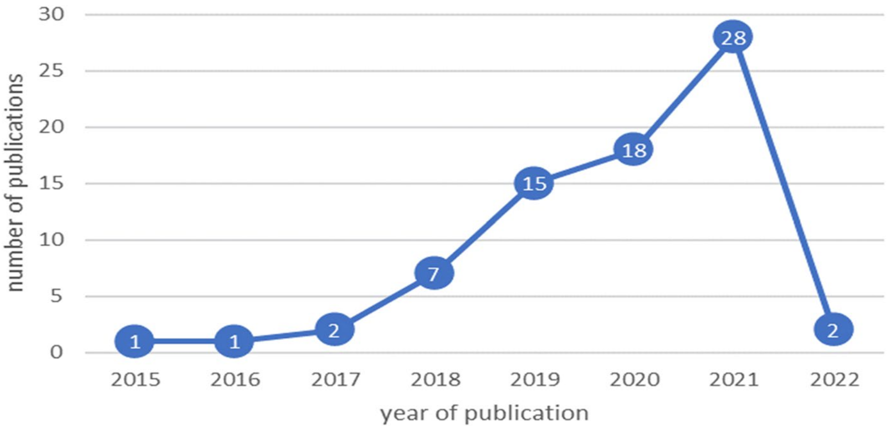

Fig. 3 Evolution of publications per year

removed. The remaining studies were then evaluated based on their abstracts. Finally, the full text of publications that were retained after the first two rounds were downloaded and thoroughly reviewed. In the end, papers that solely developed prediction models for learning outcomes and evaluated their predictability were retained. We obtained 74 studies that are analyzed in the remainder of this paper; the details of the study selection process considered in this work are presented in Fig. 2.

3.4 Data extraction

To answer the three research questions, we extract the following features to obtain relevant information for our review: (1) bibliometric data, (2) the learning outcome(s) being predicted, (3) the features extracted for use in the prediction (predictor variables), (4) the ML methods used in prediction, and (5) the evaluation measures of the prediction. We examined each individual paper that was kept following the filtering process and then identified data related to each of the aforementioned aspects. The data was aggregated and analyzed, and the findings are shown in the next section.

4 Result

In this section, we present the results of the thorough review of selected publications according to the three research questions.

4.1 Bibliometric analysis

Temporal and geographical trends As shown in Fig. 3, the 74 publications collected were published after 2015. All the literature published between 2012–2014 was

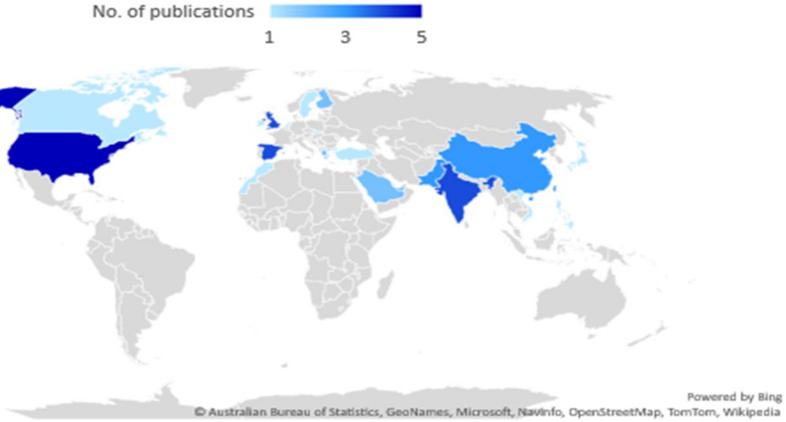
No. of publications
1
3
5
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom, Wikipedia

Fig. 4 Breakdown of the included studies by country

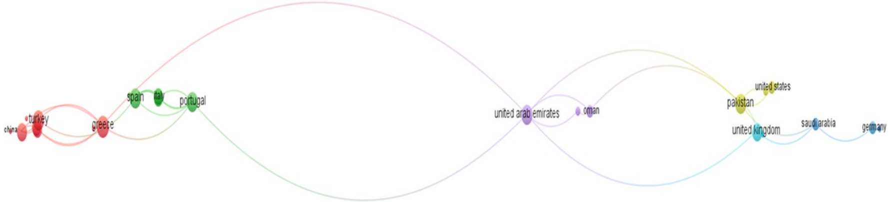

Fig. 5 Countries' collaborations

excluded, since it does not meet our inclusion criteria that require the use of at least one ML method and impose an explicit description of predictive modelling process. This shows that it took some time for ML and DNN to have a broad impact on the PLA research domain. Furthermore, we noted an upward trend for publications in the last three years. 85% of the reviewed studies have been published since 2019. 2021 is marked by the highest number of publications in the field with 28 papers – 38% of the total publications-. This shows the growing interest in PLA over the last years. Only publications made in early 2022 were included, therefore a decrease in the number of publications can be observed in 2022. However, given the extent of work and scientific endeavor in the field so far, it is expected that the upward trend will continue this year and, for the years to come.

There are 23 countries that participated fully in 63% of the total number of publications while the remaining 37% was a result of collaborations across countries.

Figure 4 represents the contribution of the different countries. Although researchers worldwide have been studying ML and predictive models in LA, the leading-edge work is mainly performed in the United States, India, Spain, and The United Kingdom. The rest of the studies are distributed between Europe (8 studies) and other Asian countries (14 studies) and the Arab world (7 studies).

In terms of collaborations, Fig. 5 illustrates the countries' co-authorship network graph. The countries with the highest total links, represented with the biggest nodes

Table 3 Types of the literature

Type of publication	No. of publications	No. of citations*
Conference Proceeding	5 (7%)	70
Journal Article	69 (93%)	2028
Total	74	2098

*The number of citations as of March 8th, 2022

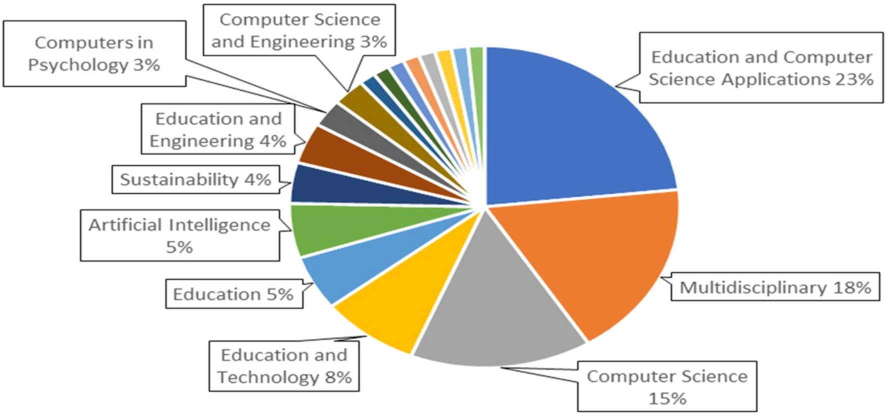

Fig. 6 Subject areas of the publications

in the graph, are the United Arab Emirates and Greece with seven collaborations, followed by Spain, Pakistan, Oman, Turkey, and Portugal each with six.

Publications' profiling The majority of the 74 retrieved publications are articles in journals with a percentage of 93.24% while the remaining 6.76% are papers in conference proceedings. This shows the maturity of the studied field. Table 3 gives the distribution of publications by their type.

Regarding the subject area of the publications, we found that "Education" and "Computer Science" are the prevalent areas representing 58% of the publications. 23% of the studies were published in the category of "Education and Computer Science Applications", 15% in "Computer Science" area, followed by 8% and 5% in the categories of "Education and Technology" and "Education" respectively. A total of 7% were found in sources combining "Education" or "Computer Science" with "Engineering". An important percentage of 18% were issued in "Multidisciplinary" sources. The remaining percentage of studies were published in other areas. Figure 6 details the distribution of publications by subject area.

Authorship analysis As follows from Fig. 7, we can observe that there is a lot of co-authorship in this field. The average number of authors is four, more than half of the publications (63%) had two to four authors. The highest number of authors per

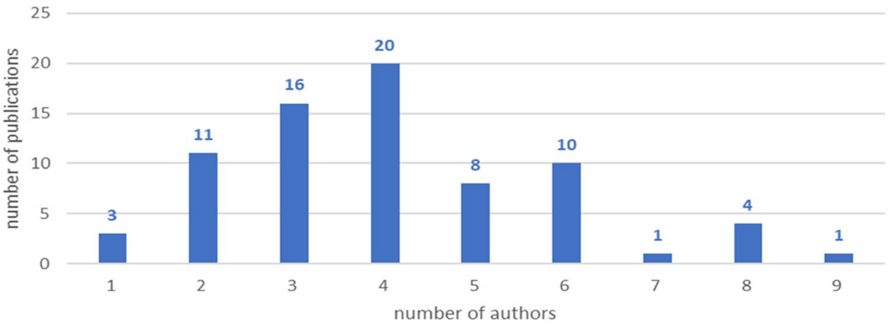

Fig. 7 Frequency of publications by number of authors

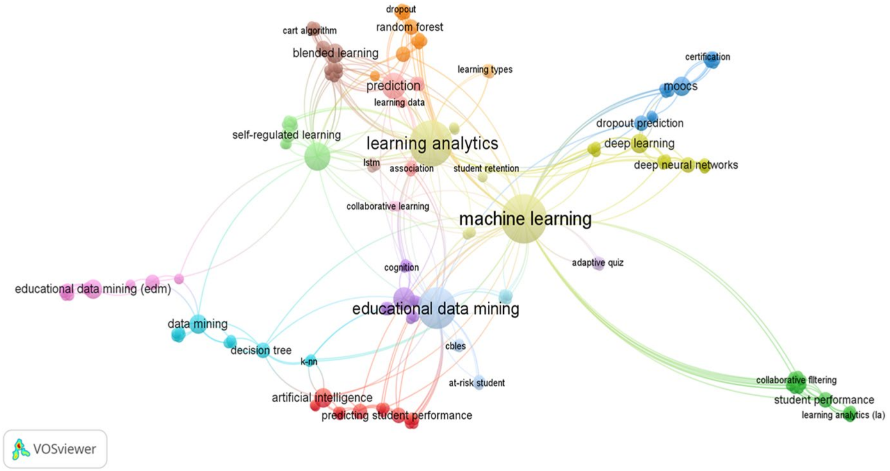

Fig. 8 Keywords network of publications

publication is nine authors in one publication, while there are only three publications with one author.

The included studies have more than 200 different authors. Most authors have only made one scientific contribution. While nine authors have contributed to this field with two or more publications.

Keywords analysis There are 172 different keywords provided by the authors, the most frequently used terms are “machine learning”, “learning analytics”, and “educational data mining”, with 17 occurrences, 15 occurrences, and 12 occurrences, respectively. Meanwhile, the term “prediction” appears 5 times on its own and 17 other times in conjunction with words such as success, performance, dropout, achievement, and at risk. Figure 8 presents the network graph of the keywords. As can be observed, the most used terms are placed in the center of the graph and have the strongest connections with the other keywords.

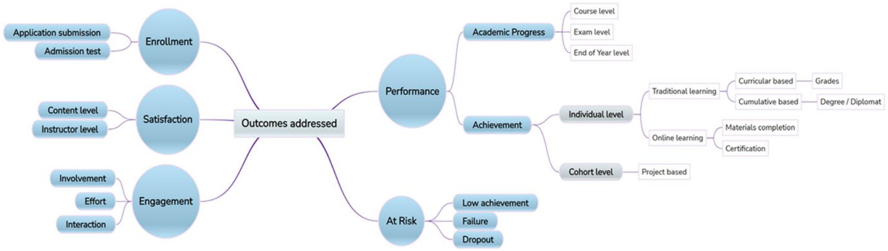

Fig. 9 Taxonomy of outcomes addressed using PLA

Seven different clusters can be identified from the keywords graph. Each cluster indicates a set of keywords that were frequently mentioned together. We noticed that the biggest nodes in the graph are linked to nodes from all of the different clusters implying that there is a significant overlap between predictive analysis and the fields of LA, EDM, and ML. The keywords related to Learning Analytics and ML are naturally located in the center of the word map, and three sub-domains are found to be clustered around these two terms. The cluster of ML includes methods that are mainly used in prediction such as: deep learning, random forest, etc. The cluster of learning analytics contains keywords related to EDM, e-learning, and collaborative learning. And the last sub-domain represents keywords related to learning outcomes such as dropout, performance, at risk and so on.

4.2 Outcomes and inputs analysis

The studies analyzed in this review vary mostly in terms of the issue being addressed or predicted (the outcome variable) and the factors or measures that exert a significant effect on this issue (predictor variables). In this section, we review the most frequent issues that PLA is used to address and the main features that are proven to be correlated with. To begin with, we discuss the learning outcomes and compile them into a taxonomy. Then, we go through the main features that were used. Finally, we draw links between outcomes and features to map the relationships built by predictive models in the current literature.

Detailed examination of the scanned body of literature showed how PLA can be used to tackle diverse learning problems; it can be exploited throughout students' whole academic careers starting from their enrollment to graduation. Specifically, current contributions in the literature can be categorized into five main classes, namely (i) Enrollment, (ii) Performance, (iii) At risk students, (iv) Engagement, and (v) Satisfaction. Figure 9 shows a visual representation of the different classes of the predicted outcomes in the literature.

(a) **Enrollment**: predictive models can improve the process of enrolling by analyzing students' data to assist them in their orientation thus improving their chances

Table 4 Frequency and impact of the predicted outcomes by the number of publications and citations

	Nb of publications	Nb of citations*
Performance	57	1518
At risk	20	714
Students' Engagement	5	254
Satisfaction	4	169
Enrollment	3	22

*The number of citations as of March 8th, 2022

of acceptance. Studies addressing enrollment focus either on the admission test, the submission application, or both.

(b) **Performance:** the overall success of an education system can be measured by the academic success of its students. Predictive models play an important role in the early evaluation of students' learning outcomes and the assessments of their performance. Indeed, reviewed literature tackles performance from different points of view. Specifically, the performance analysis can target (i) the continual academic progress at the level of the course, assignments, or annual final examination. It can also target (ii) the overall achievement at the individual level or collective level. Individual performance analysis is further investigated based on traditional learning elements which include curricular-based analysis (e.g., grades) and cumulative-based analysis (e.g., diploma), or online learning aspects including e-certifications and material completion.

(c) **At risk students**: by the use of predictive modelling methods, it is possible to identify at-risk students early and implement preventive and corrective measures to increase the students' commitment to their educational goals and boost the rate of retention. The scanned literature identifies at risk students through (i) low achievement, (ii) failure, and (iii) risk of dropout.

(d) **Engagement**: another use of PLA is surveying the learning behavior of students and their level of engagement to adapt curriculum, ameliorate their contents, and change the style of teaching to support their success. The works that fall in this class analyze mainly behaviors related to (i) involvement, (ii) deployed efforts and (iv) interaction.

(e) **Satisfaction:** the last class of studies investigates how students perceive the services they are offered and how satisfied they are with them. The satisfaction is measured at (i) the content level and (ii) the instructor level.

It is important to note that the classes described in the proposed taxonomy are not meant to be mutually exclusive. Indeed, overlaps can be observed between the identified classes. Some works can belong to both engagement and performance studies for example. The proposed categorization should be considered as an abstraction of the present state of thinking that can be useful for interested researchers to learn about the main issues shaping the landscape around PLA.

Naturally, the identified outcomes have not been equally addressed in the literature. As depicted in Table 4, there is a variation in the frequency of the outcomes in the studied papers. The most frequent aim of using PLA in higher education is the student performance assessments, thus enhancing their learning outcomes. Next comes identifying students at risk, which also received considerable attention from researchers. Providing insights on patterns that increase the success rate and reduce the failure of students is profoundly helpful to educational institutions, this can explain the focus on these two outcomes. On the other hand, the little regard that has been given to the other outcomes may be explained by the fact that the study of engagement and satisfaction involves complex behavioral analysis. Moreover, these types of outcomes are often viewed as implicit (indirect) influencers on student success.

In the context of performance, several authors (Afzaal et al., 2021; Albalooshi et al., 2019; Chen & Cui, 2020; Gitinabard et al., 2019; Jensen et al., 2021; Mai et al., 2022) predicted learner's performance in a course level by targeting their scores in quizzes, assignments, or exams. Other publications focused on the degree level (Adekitan & Salau, 2020; El Aouifi et al., 2021; Tuononen & Parpala, 2021; Villagrá-Arnedo et al., 2016), while Zeineddine et al. (2021) predicted performance of new-start students based on first semester results. They evaluated students' graduation using cumulative indicators to forecast their final grades. Aside from individual achievement, Ekuban et al. (2021) and Spikol et al. (2018) projected performance from a cohort viewpoint based on group projects and team success.

In addition to performance prediction, efforts were made to identify at-risk students to assist them and implement intervention programs to match those students. For instance, works described in (Albreiki et al., 2021; Karalar et al., 2021; Macarini et al., 2019; Rafique et al., 2021; Zacharis, 2018) predicted which students are most likely to fail in the early stages. Another research described by Gray and Perkins (2019) and Chui et al. (2020) identified marginal students who are at-risk of low achievement. Authors in (Adnan et al., 2021; Dass et al., 2021; Goel & Goyal, 2020; Kabathova & Drlik, 2021) identified dropout students based on their academic results.

Aside from educational outcomes, only a few studies analyzed other types of outcomes related to learners' behavior and emotion. Studies detailed in (Ayouni et al., 2021; Dias et al., 2020; Hsu Wang, 2019; Hussain et al., 2018) predicted students' engagement based on their involvement and interaction with course materials, another research mentioned by Sharma et al. (2019) targeted their effort. Heilala et al. (2020) and Hew et al. (2020) on the other hand, predicted students' satisfaction with the course content and the instructor. Furthermore, predictive analytics was employed by Moreno-Marcos et al. (2019) to forecast admission test outcomes while Iatrellis et al. (2021) and Sghir et al. (2022) predicted the likelihood stream of enrollment.

In LA, predictive models use many predictor variables as inputs depending on the type of prediction insight. Similarly, to outcome variables, we also classified the common features in the extracted works into three main categories that cover the whole educational ecosystem, namely: (i) student-related, (ii) teacher-related, and (iii) institutional features.

Starting with the student-related category, we distinguished between five types of data which are:

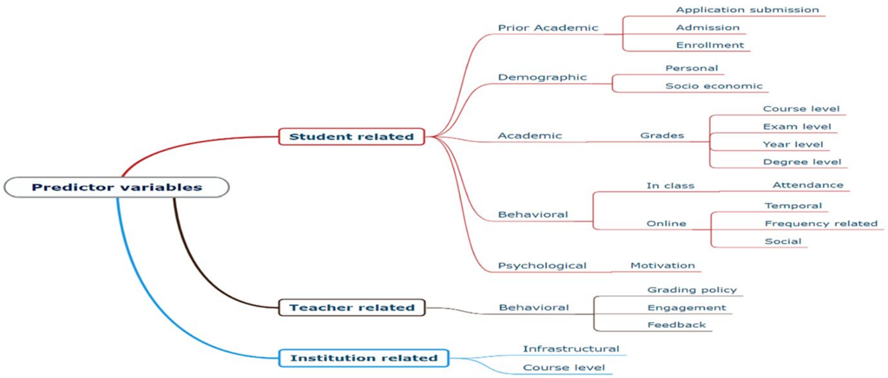

Fig. 10 Classification of features used in the literature using predictive modelling

- Prior academic data containing students' historical records in previous studies, and their admission information,
- Demographic features covering personal and socio-economic characteristics,
- Academic data represented by students' achievements during their higher education,
- Behavioral features extracted from learners' actions and activities,
- Psychological data compromising students' motivation, interest in the course, or trust in the instructor.

In the teacher-related class, only behavioral data was employed in the prediction. However, in the institutional-related category, we differentiated between two types of variables, the first type was related to the institutions' infrastructure and the other one covered the course content. A detailed classification of the input features is described in Fig. 10.

In addition to identifying the main outcomes and input features that are used in the existing predictive learning models, it is more crucial to get insights into the relationship between inputs and outcomes. Based on the extensive review of the findings of the selected studies, we elaborate a correlation map that is described in Fig. 11 and detailed further in the following sub-section.

Most of the recent studies employed students' online behavioral data in the prediction since most of the experiments were conducted in an online learning environment or a blended setting. Some of the examples are presented in (Abdullah et al., 2021; Cerezo et al., 2017; Lu et al., 2018; Mansouri et al., 2021; Shayan & van Zaanen, 2019), authors used students' behaviors logs containing their interactions with the online content to assess their performance, while Ayouni et al. (2021) and Hussain et al. (2018) used it to measure students' levels of engagement. (Chen & Cui, 2020; Chen et al., 2020; Dass et al., 2021; Macarini et al., 2019) utilized features related to time spent on online materials to identify at-risk of failing students or the ones who are most likely to dropout. However, other authors (Adnan et al., 2021; Goel & Goyal, 2020; Karalar et al., 2021; Waheed et al., 2020; Yu et al., 2019) utilized behavioral data compromised in students' clickstreams data. Another

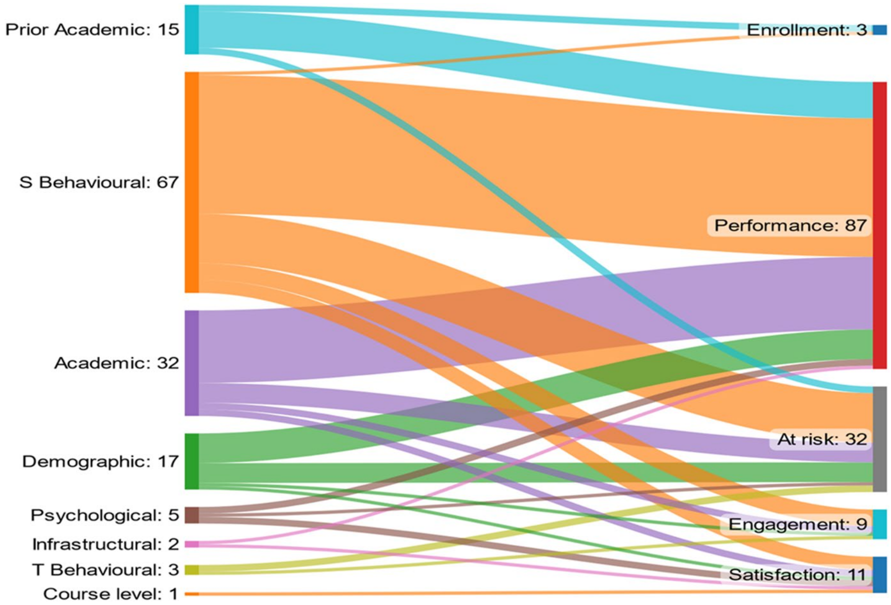

Fig. 11 Correlation between the predictor variables and the predicted outcomes

example may be seen in Joksimović et al. (2015) and Zacharis (2018), where the authors focused on characteristics associated with learners' social participation in online communities to predict students' achievement and classify the low achievers. Meanwhile, Gitinabard et al. (2019) employed features extracted from learners' study habits and their social interactions to assess their performance. In traditional learning settings, the in-class type, (Gray & Perkins, 2019; Khan et al., 2021; Parvathi, 2021) employed students' attendance behavior as an indicator to predict their learning outcomes.

In addition to behavioral data, students' academic data is the second most commonly used type of features. Several studies used learners' grades in a course, exam, year, or degree level as an academic indicator to predict different outcomes. An example is demonstrated by Mai et al. (2022) where the authors used students' assessment scores in an online learning system to predict the overall mark in the terminal exam. Whereas Albreiki et al. (2021) employed academic data containing homework assignments, quizzes, mid-term exams, and final exam marks to identify students at-risk of failing. A different use of academic data was to measure students' level of engagement (Ayouni et al., 2021).

Next comes students' demographic data as predictor variables which were mostly used for academic outcomes. Prada et al. (2020) used students' gender, age of access to studies, nationality, parents' education level, and residence as part of the inputs to measure students' performance on a degree level. Another example was described by Rafique et al. (2021) where the authors employed students' age, gender, residence, location, and father's qualification and financial status to highlight at-risk students during the early weeks of a course and to uplift the performance of low-performing students. Different usage of demographic data was to determine students' level of engagement in an e-learning system (Hussain et al., 2018).

In most cases, prior academic type of data was employed in the performance prediction. Almasri et al. (2019) used attributes collected from students' historical educational records, while Adekitan and Salau (2020) used students' preadmission scores to predict their graduation grades. In a few other cases, as reported by Rafique et al. (2021), prior academic data was used to identify at-risk of dropping out utilizing features extracted from previous studies and access scores. It may also be employed in enrollment prediction (Iatrellis et al., 2021; Sghir et al., 2022).

Students' psychological data is not commonly employed in predictive modelling, only a few publications have documented its usage. An example is mentioned by Heilala et al. (2020) where authors relied on questionnaire responses to collect psychological features such as their trust and interest to measure their course satisfaction.

Aside from the student-related category, there has been relatively little literature published on the use of features related to the teacher or institution. Regarding the teacher-related category, Lu et al. (2021) used features associated with the teacher grading style as inputs to predict students at-risk of failing, whereas Herodotou et al. (2019) employed data compromised in the instructor's interactions with an online learning environment to identify students with a high probability of not submitting or failing the next assignment. Ayouni et al. (2021) on the other hand, utilized teachers' appreciation as a predictor variable to assess students' levels of engagement.

As for the institutional factors, Hew et al. (2020) utilized features related to the course content to predict students' level of contentment. However, Mansouri et al. (2021) used the quality of the institution's infrastructure as a part of the inputs to evaluate students' performance and measure their satisfaction.

In the extracted studies, a combination of multiple types of features was employed simultaneously depending on the outcome. Figure 11 gives a graphical representation of the relation between the predictor variables and the predicted outcomes.

As can be seen from Fig. 11, most of the studies used behavioral data related to students to measure their performance primarily and identify the ones with a probability of failing or dropping out. Students' academic data comes second, with the majority of it being used to predict students' performance and the ones at-risk, while the remainder was used to measure students' levels of engagement and satisfaction. In the prediction of enrollment, students' prior academic data were used as predictors for the stream of enrollment. Meanwhile, students' behavioral data was employed to forecast the results of the admission test.

Aside from the correlation displayed in the figure above between the categories of the inputs and the predicted outputs, we observed that certain predictors have a greater impact on students' outcomes than others. In the behavioral data, for example, features related to time management influence students' achievement positively. Students who devote enough time to their academic work and complete it on time or early demonstrate satisfactory performance. Prior academic and academic data; including admission score, previous grades, and early performance, affect students' graduation positively and can help predict those at risk. Psychological data, such as students' motivation, can also have an impact on their overall performance.

However, the impact of predictor variables on student outcomes varies from one context to another.

4.3 PLA in response to COVID-19 pandemic

PLA has a pivotal role to play in informing stakeholders of the higher education system about the impact of disruptive events such as the pandemic. Indeed, the move to online learning during the prevalence of COVID-19 has yielded access to a large amount of educational data, on one hand. On the other hand, the loss of direct student contact has caused frustration among educators not knowing whether or how students were learning and engaging. This situation provided an unprecedented opportunity for PLA to show how AI and ML can help in gaining insights into student progress, needs, potential, and risk during a crisis or major changes. In the analyzed literature, the studies of (Abdullah et al., 2021; Dias et al., 2020) were found to specifically mention or address the role of PLA in the COVID-19 context. For instance, Dias et al. (2020) assessed students online learning interactions to predict the quality of their engagement during the pandemic. This modest extent of literature illustrating the response of PLA to the pandemic is expected to expand. It is likely that further findings and more detailed studies of the changes in PLA practice in the post pandemic era will emerge over time. Technically, the potential of PLA to manage situations caused by the pandemic is hindered by the so-called data drift (Adadi et al., 2021). In order to deliver prediction, models analyze historical data. However, pre-COVID data are no longer relevant since they do not include new student and teacher behaviors and interactions. For example, most of the predictions of the 2020–2021 admissions cycle were probably inaccurate, since the models based their prediction on historical data that does not reflect the impact of a global pandemic on student admissions and retention. ML models degrade gradually when the data they were trained on no longer reflect the present state of the world (Adadi et al., 2021) (i.e., data drift). However, opening up to online learning environments accelerated by the pandemic can help in collecting new data that accurately reflect the new learning behaviors.

4.4 Predictive modelling analysis

In this segment, we will go over the methods utilized to predict student outcomes presented in Table 5. We will discuss the predictive analytics process and the most often employed approaches at each step.

To forecast a learning outcome, the first step is to identify the type of prediction. In the reviewed studies, the prediction problem was approached from two different angles using:

1. Classification: predicting whether the desired output belongs to a category or another.
2. Regression: predicting the continuous value of the target outcome.

Table 5 The main used learning algorithms

Task	Learning algorithm	Nb of articles
Classification	ANN	10
	RF	7
	GB	5
	DT	4
	NB	3
	Ensemble ML	3
	KNN	2
Regression	MLR	4
	Linear R	2
Clustering	K means	7

In order to build a predictive model capable of accurately predicting the desired outcome. Researchers followed the steps described below.

4.4.1 Data collection

In predictive analytics, models initially learn from learners' past data to anticipate their future behaviors. It is of the utmost importance to obtain reliable data to create accurate predictions. The collected data varies in terms of volume and sources. Researchers followed several strategies to collect data in the studied literature, such as:

a) Learning platforms; as mentioned by Dass et al. (2021) where data from 3172 students was derived from a self-paced math course College Algebra and Problem Solving offered on the MOOC platform Open edX by EdPlus at Arizona State University (ASU). Lincke et al. (2021) extracted logs of 300 medical students through an adaptive web-based learning platform used in Sweden named Hypocampus. Another example is mentioned by Cerezo et al. (2017), the interactions of 140 undergraduate psychology students were gathered from a Moodle program named eTraining for Autonomous Learning in a state university in Northern Spain. Meanwhile, Emerson et al. (2020) obtained data from 61 College students from three large North American universities through Crystal Island, a game-based learning environment for microbiology education.
b) A second source of data was Student Information Systems; as stated by Hasan et al. (2020), 772 students' academic information were collected from a private institution's system in Oman. Meanwhile, Cutad and Gerardo (2019) used a dataset including 343 students records from the Enrollment Information System of the Information Technology Students at Davao Del Norte State College in the Philippines.
c) Questionnaires are another data source; as an example, Tuononen and Parpala (2021) employed an electronic questionnaire to gather responses for analysis from 1019 students who had completed their bachelor's or master's degrees from the

Faculty of Arts and the Faculty of Social Sciences in Finland. While Rafique et al. (2021) obtained data from 164 first-year undergraduate students at COMSATS University's Computer Science Department using a questionnaire that included questions about their personal, socioeconomic, psychological, and academic information.

d) Public data sources are a simple way to acquire data for analysis. For instance, (Adnan et al., 2021; Chui et al., 2020; Waheed et al., 2020) utilized the Open University Learning Analytics Dataset (OULAD), a freely available dataset of 32,593 students provided by the Open University in the United Kingdom.

e) A residual percentage of studies employed multi-modal data; for example, Sharma et al. (2019) used 32 undergraduate students from a European University for the experiment and collected their sensor data from four different sources: eye-tracking, EEG, facial video, and arousal data from the wristband. Another example is described by Spikol et al. (2018), in which the extracted data from 18 engineering students included the capture of objects, the positions of people, hand movements, faces, and audio levels and video, as well as interactions of plugged components from the Arduino-based physical computing platform and the interaction with the sentiment buttons.

Regardless of the various data collection techniques mentioned above, we discovered that learning platforms are the most widely used data sources due to the increased adoption of blended learning settings and fully online environments by institutions.

4.4.2 Data pre-processing

This step is critical to ensure that raw data is consistent, clean, and can be useful. Real-world data is usually inconsistent, incomplete, or may contain inaccuracies. To address such problems, different techniques are used for pre-processing data to improve the predictive power of models.

- Data cleaning: is the process of eliminating, altering, or formatting data that is incorrect, irrelevant, or duplicated because it can offer misleading or incorrect insights. In the reviewed literature, missing variables instances were removed (Afzaal et al., 2021), replaced by global constants like zero values (Hussain et al., 2018), substituted by their mean values (Adnan et al., 2021), or estimated using the k-nearest neighbor imputation (Al-Shabandar et al., 2017).
- Feature engineering: it entails either transforming existing features or generating new features using the original ones. Feature encoding is used to convert categorical data types into numeric vectors that can be interpreted by a learning algorithm. Researchers employed techniques such as label encoding (Goel & Goyal, 2020) and one hot encoding (Rafique et al., 2021). Another step is feature scaling, which involves bringing all the features to a common scale, it ensures that a feature with relatively larger values will not control the model. In

the extracted studies, normalization (Iatrellis et al., 2021), and standardization (Chen & Cui, 2020) were the most utilized techniques for scaling.

- Feature selection: consists of reducing the input data by identifying the best set of independent features to help improve the prediction model's performance. Various techniques are used including the Correlation matrix to understand the dependency between the features and the targeted learning outcome (Kabathova & Drlik, 2021), the Information Gain (Khan et al., 2021), and the Gini coefficient. Recursive Feature Elimination (RFE) to remove the weakest features (Chen & Cui, 2020), the Genetic Algorithm (GA) (Hussain & Khan, 2021), and finally the Sequential Forward Selection (SFS) by selecting the strongest features (Gray & Perkins, 2019).
- In the case of an imbalanced dataset where the frequency of some classes is very high in comparison to others. Researchers used different sampling techniques to overcome the problem. The most used one was over sampling by augmenting the minority class. Macarini et al. (2019) applied Synthetic Minority Oversampling Technique (SMOTE). While Rafique et al. (2021) employed an adaptive synthetic (ADASYN) over-sampling approach to resolve the dataset imbalance problem.

4.4.3 Learning algorithms

The main aim of this step is to choose suitable algorithms for predictive analysis based on several characteristics such as the type of the problem, the size of the dataset, the nature of the input data, and the outcome to predict. In the analyzed literature, researchers tended to utilize more than one algorithm to select the best model for prediction depending on its performance. Examination of the studies showed that Artificial Neural Networks (ANN) attained the best performance and are thus the most used in prediction, followed by Random Forest (RF) and Gradient Boosting (GB) algorithm. Table 5 shows the distribution of learning algorithms based on the number of articles. We considered algorithms that appeared in more than one publication.

ANN have played an essential role in predicting learning outcomes, they can solve a variety of issues and predict whatever targets are required. As an example, Neha et al. (2021) proposed a deep neural network for evaluating student performance assessments, meanwhile Ayouni et al. (2021) used an ANN to predict students' engagement in an online environment. In (Chen et al., 2020), researchers employed a hybridized deep neural network to identify students at risk early in the exams. (Chen & Cui, 2020; Dias et al., 2020; Mubarak et al., 2021) utilized the recurrent neural network named long short-term memory (LSTM) to predict course performance, evaluate the quality of interactions and students' involvement, and predict students' weekly performance, respectively. A different approach was presented by Abdullah et al. (2021), where a feed forward neural network was trained to assess students' academic performance.

The second most used algorithm in prediction was RF. Dass et al., (2021) and Kabathova and Drlik (2021) utilized the algorithm to identify students at risk of dropping out of an online course. Al-Shabandar et al. (2017) and Hasan et al. (2020) employed the model to predict students' performance. However, Heilala et al. (2020) utilized it to measure course satisfaction, while Ekuban et al. (2021) applied the model to predict students' group success. Iatrellis et al. (2021) developed tow RF models to predict graduation and students' enrollment in postgraduate studies.

Tree-based models, including algorithms such as Decision Tree (DT) and GB, are other well-known prediction methods. For instance, Lincke et al. (2021) compared seven algorithms including linear and logistic regressions, gradient-boosted tree, XGBoost, deep neural network, Bayesian neural network, and rich context model (RCM). The models were applied to a medical student dataset for quiz performance prediction. The authors found XGBoost as the model with the highest performance. However, RCM was found to be more transparent, and the reasoning behind its result is easy to explain. Other examples were presented in (Hussain & Khan, 2021; Mai et al., 2022), the studies selected GB, and DT, respectively, to forecast students' overall performance. Kostopoulos et al. (2021) employed GB algorithm to predict certification in an online course, meanwhile Hew et al. (2020) used the same algorithm to measure students' level of satisfaction.

In terms of regression problems, Single and Multiple Linear Regression (MLR) were used for prediction. Hsu Wang (2019) employed the multiple linear regression model to predict students' online behavior and their achievement. Meanwhile, it was used by Albalooshi et al. (2019) and Yang et al. (2018) to assess learners' academic performance. Single linear regression was utilized by Omer et al. (2020) to evaluate the performance and by Tuononen and Parpala (2021) to predict students' thesis grades.

Naive Bayes (NB) followed closely. The algorithm based on the Bayes theorem was employed to predict at risk students (Bañeres et al., 2020), assess students' academic performance (Alturki et al., 2021), and estimate success in a project-based learning setting (Spikol et al., 2018).

To predict learners' performance, K Nearest Neighbor (K-NN) algorithm was employed (El Aouifi et al., 2021), while it was used by Cutad and Gerardo (2019) for curriculum analysis.

In other cases, researchers combined different base models to make a more accurate prediction. An ensemble-based model combining Extra Trees, Random Forest, and Logistic Regression was employed by Karalar et al. (2021) to predict at risk of failing students. Meanwhile, the ensemble model (Rafique et al., 2021) combined Support Vector Machine (SVM), RF, and K-NN to evaluate academic performance. Zeineddine et al. (2021) presented another approach in which an automated ML (AutoML) process was used to improve the accuracy of predicting student performance at the start of their first year.

After choosing the learning algorithms, designing a good training and testing procedure is crucial to determine whether a predictive model is accurate when exposed to new data in a real-world environment. Cross Validation was used by researchers as a way of measuring the generalization of the model. The most used techniques in the studied publications were K-fold cross validation, followed by the Leave-One-Out cross validation method.

Table 6 The main used performance metrics

Task	Metric	Nb of articles	
Classification	Accuracy	43	85%
	F Measure	22	
	Recall	19	
	Precision	18	
	AUC—ROC	16	
	Kappa	6	
	Sensitivity	4	
	Specificity	4	
	MCC	2	
Regression	Pearson R	5	15%
	RMSE	4	
	pMSE	2	
	pMAPC	2	

4.5 Evaluation metrics analysis

Considering that the output of predictive modelling is probabilistic, evaluating the obtained results is a crucial step. There are a lot of ways to judge how well the model performs to make appropriate changes and, as a result, produce more accurate predictions. The evaluation metrics are determined by the type of the problem. In classification problems, the authors of the selected studies used the confusion matrix and the measurements drawn from it. Meanwhile, for the regression cases, they used metrics like Root Mean Squared Error (RMSE), Mean Absolute Error, and R-Squared. Different evaluation metrics were used in the same study to further calculate the models' performance, Table 6 describes the distribution of the metrics used by the type of prediction problem.

Concerning the evaluation of classification problems, an example is mentioned by Kabathova and Drlik (2021), where the Random Forest model obtained the highest precision, F1 scores, and accuracy. The precision of the model reached 86%, the recall was 96%, while the F1 score was 91% and with an overall accuracy of 93% to predict students' dropout in course level. Meanwhile, the same model predicted the dropping out or continuation of students on any given day in a MOOC course with an accuracy of 87.5% (Dass et al., 2021). In a separate study (Lincke et al., 2021), the best models in predicting the probability that a student will answer correctly in a quiz were gradient-boosted tree and XGBoost with around 88% accuracy with AUC values of 0.903, and 0.94 respectively. Since the ROC curve is more applicable for balanced datasets, the authors used another metric called Precision-Recall to judge their models due to the unequal distribution of classes in their dataset. However, to assess students' performance, Khan et al. (2021) employed the metrics of accuracy, precision, recall, and F-Measure were not enough to select the best model. The decision tree model achieved the highest F-Measure of 0.91 and an accuracy of over 85% but it scored a lower recall value in comparison to the other models. To check the ability of the models, the authors used another metric called Mathew Correlation Coefficient (MCC), the metric corroborates the decision tree pre-eminence with an MCC of 0.63. Prada et al. (2020) developed a tool called SPEET, the evaluation of its performance was determined by the accuracy, sensitivity, and specificity. The proposed model predicted dropping out with an accuracy that reached 90.9%, a sensitivity of 97%, and a specificity of 75.4%. To predict at-risk students, Waheed et al. (2020) noted that the deep artificial neural network model outperforms the traditional algorithms with an accuracy of 84%-93%. Another deep learning approach was employed by Chen and Cui (2020) for the early prediction of course performance, the study used a recurrent neural network named LSTM that achieved an AUC above 70% in comparison to other methods. Meanwhile, Abdullah et al. (2021) employed a feed forward neural network classifier that achieved an accuracy of 88.18%. A different approach was mentioned by Rafique et al. (2021), the authors used ensemble-based model to predict at risk students, the model combines different base models such as SVM, RF, and K-NN as sub-estimators, it showed better performance than individual models with an accuracy of 82.98%.

To evaluate prediction performance in regression problems, different metrics were used. As an example, Jensen et al. (2021) trained a RF regression model to predict a student's quiz score, the model obtained a high accuracy with Pearson r = 0.53. R-squared was utilized by Hsu Wang (2019) to evaluate the predictive model, the multiple linear regression algorithm achieved an R-squared value of 0.511. Another metric is presented by Mansouri et al. (2021), where researchers employed mean squared error (MSE) and Standard deviation to assess the performance of a new approach based on the Learning Fuzzy Cognitive Map (LFCM). However, Yang et al. (2018) applied the predictive mean squared error (pMSE) and predictive mean absolute percentage correction (pMAPC) as measures. The MLR used to forecast students' academic performance achieved an optimal pMSE of 198.62 and a pMAPC value of 0.81, it accurately predicted the academic scores of 8 out of 10 students. Meanwhile, Lu et al. (2018) employed principal component regression that obtained pMSE and pMAPC values of 159.17 and 0.82, respectively.

5 Discussion

Examining the reviewed literature provided a comprehensive overview of the adoption of PLA in higher education over the last decade. The selected studies contained qualitative and quantitative information about the predictive modelling and the key elements to implement it, as well as the benefits regarding its use in higher education.

To answer the first research question (RQ1), we conducted a bibliometric analysis of the scanned body of literature, the results showed that PLA is a very dynamic research area, worth following in the coming years. As shown in the greater number of publications and citations over time with a higher proportion of journal articles. The more countries/regions implication in the related research, and the study of various educational problems/outcomes.

In general, PLA focused on the discovery of aspects related to learning processes and, as a result, learning outcomes. It could bring major educational benefits such as enhancing learning results, early detection of students with a high probability of failing or dropping out, increasing learners' level of engagement, and raising their satisfaction with the learning process. Concerning the second research question (RQ2), our study revealed that predictive modelling was mostly used in assessing students' performance and detecting those at risk to intervene and help them to ensure their success and boost the rate of retention. While satisfaction and enrollment related outcomes have been addressed less frequently in the studied literature. The lack of attention paid to these outcomes may be attributed to the fact that they do not have a direct influence on student success in comparison to other outcomes, or maybe it is still difficult to measure and quantify a complex, emotional, and subjective phenomenon such as satisfaction. During the pandemic and beyond, PLA was mainly deployed for supporting online learning to assess students' interactions and evaluate their e-mental health to help them overcome psychological distress. Data drift is an issue to overcome in future research in order to produce accurate predictions that consider new learning behaviors.

Furthermore, to predict a learning outcome, researchers employed a variety of predictor variables, we grouped the extracted variables into three main classes: student-related, teacher-related, and institutional features. Due to the rising usage of online learning settings especially after the outbreak of the COVID-19 pandemic, students' behavioral data, particularly their online interactions, were the most commonly employed as inputs for prediction. These data logs offer the possibility of predicting different learning outcomes through the patterns extracted from students' behavior. Based on our correlation analysis, the relation between behavioral data and student performance is the most studied in the reviewed literature.

The predictive modelling process involves several steps that include determining the best combination of tools and methods to obtain the optimal result in the prediction process. One critical step is the selection of a learning algorithm to achieve accurate predictions, the choice depends on different aspects starting from the type of the problem to the nature of the outcome to predict. To answer the third question (RQ3), we conducted a thorough examination of the predictive models used by the reviewed studies. Back in 2014, Baker and Inventado (2014) noted that although prominent in other data mining domains, neural networks are somewhat less common in EDM. They believe that the complexity of the educational domains leads the community to choose more conservative algorithms. Some previous works even advised educational researchers to favor the explanation provided by classical ML over the accuracy provided by DNN as criteria for selecting computational techniques for LA/EDM research (Doleck et al., 2020). However, a few years after facts have changed. Now, neural networks especially their current incarnation, Deep Learning, are the method of choice for predictive tasks in LA and EDM. Indeed, our review showed that predictive analysis based on Artificial Neural Networks is predominant and achieved high accuracy in comparison to other models. However, it also has higher computational complexity and hardware performance needs as well as being black box models. Traditional ML algorithms, namely Random Forest and Gradient Boosting came in second and third, respectively, in terms of performance and usage frequency. To select an algorithm, researchers tended to use multiple models to determine the best one for prediction based on its performance. In the scanned studies, the performance of algorithms was evaluated using different metrics and the most employed ones were the confusion matrix and the measurements derived from it, the Mean Squared Error, and R-Squared coefficient.

While this review demonstrated the benefits of PLA and how it assists students through their learning process, it also noted some limitations and shortcomings that should be addressed by the research community in future works. One frequent limitation lay in the limited size of the datasets, researchers faced challenges dealing with small sample sizes, restricting their practical applications in real-world settings. Indeed, ML particularly deep neural networks are data-hungry models, they need large amounts of training data to generate accurate predictive models. However, number of publications reporting on using datasets of a small number of student groups impacts the overall performance. Another concern relates to the generalizability of the findings since the majority of the studies were conducted in a limited context. It will be necessary to delve more deeply into the predictive models' generalizability by analyzing diverse learning environments using various combinations of modalities and with varied student populations. Exploring data-efficient methods (Adadi, 2021) including Data Augmentation and Transfer Learning as well as sharing more and more public datasets will surely accelerate the research on PLA.

Another issue that should be addressed is privacy. PLA involves collecting sensitive data from students. The protection mechanisms of such data are not systematized in this LA pipeline, which leads to a negative perception of the use of LA in general. Hence, it is important to address the question of how privacy and security can be preserved when collecting and using educational data. A future step could consist of enhancing techniques for data anonymization to secure learners' sensitive data and ensure the protection of their information. In the context of a collaborative or decentralized LA process involving different institutions, it could be promising to instigate technologies such as Federated Learning (Guo et al., 2020) for privacy-preserving. Emergent technology Blockchain (Raimundo & Rosário, 2021) could also be promising for preserving the identity of students and securing their data.

One of the intrinsic drawbacks of ML models especially deep learning is the lack of interpretability, as it is difficult to determine what process has been used to determine the output. In certain cases, it is extremely difficult to understand why certain prediction was made. In consequence, educational managers may not trust such models to support their decision-making, especially if the future of a student is at stake. Explainable AI (Adadi & Berrada, 2018) is seen as a promising mechanism to increase algorithmic transparency and trustworthiness, which holds the potential to make LA outcomes more understandable and thus more acceptable. Incorporating visualization tools based on the prediction results to highlight students' learning progress, will also help the learner comprehend his or her learning advancement more easily, and it will provide information for the teacher to understand each student's capability in the best way possible.

The COVID-19 pandemic, considered as the largest global crisis in a generation, has changed the norms of higher education and opened the door for new ways of learning, schooling, and socializing. Accordingly, it is normal to see new factors starting to influence academic engagement and performance. Hence, in the future, more focus is expected on predictor variables related to student physical health and psychological state as well as student internet (virtual) behavior (Flanagan et al., 2022; Qiu et al., 2022). Specific real-time data types will be more used, such as face recognition, hand and eye tracking, movement detection, clickstream data, and biomedical signals analysis. While static data such as students' demographics and academic status will take the back seat. Outcomes variables, on the other hand, would shift towards emotional engagement and enhancement of students' satisfaction, creativity, and soft skills. Furthermore, as predictive learning will be more democratized in educational institutions, analysis of data related to other actors in the educational ecosystem including teachers, supervisors, and social workers will rise. Finally, Precision Learning as a way of adapting learning systems to meet the needs of an individual student is a big subject of interest for PLA that needs more attention from the researchers in the field. Personalized Recommender Systems and Chatbot Assistants can play an important role in this regard to supply learners and teachers with personalized suggestions, information, and eventual warning in accordance with their prediction outcomes.

Finally, it should be noted that this review has some limitations. First of all, only articles in English with full papers available were included. This could result in review bias, however, we found it difficult and infeasible to review studies in all languages. Second, we only focused on works that "explicitly" limited the scope of their study to the Higher Education context and predictive modelling field. Thus, some relevant findings included in papers studying education in general might be not analyzed in this review. Last, smart predictive modelling is undergoing continuous development. On the other hand, a systematic review follows a rigorous and long selection process. The review process began in March 2022, and since then the amount of literature on the subject had to be increased. However, given the type and diversity of the database sources used in this review and the selection process adopted, we are confident that a large part of the literature was covered and that the findings describe well the current state of research.

6 Conclusion

This paper provided a systematic review of recent studies in PLA to determine current trends and advancements in the field. Findings reveal that most of the existing publications utilized predictive modelling to assess students' performance and predict those at risk of failing or dropping out. Meanwhile, other learning outcomes related to learners' engagement, satisfaction, and enrollment have been addressed less frequently. Due to the increased use of blended/online learning settings, particularly following the outbreak of the COVID-19 pandemic, most of the students' data used for analysis came from electronic sources. This explanation also holds true for the types of features considered in publications; students' behavioral data extracted from their logs were the most employed as predictors. As for the techniques used in the prediction, Artificial Neural Networks, Random Forest, and Gradient Boosting placed first, second, and third, respectively, in terms of prediction accuracy and usage frequency in comparison to other algorithms. The performance of algorithms was commonly evaluated using the confusion matrix and the measurements obtained from it, the Mean Squared Error, and R-Squared coefficient. Spotted Limitations open up the research horizon for more innovative, data-efficient, explainable, and privacy preserving models of prediction.

Abbreviations *AI*: Artificial intelligence; *PLA*: Predictive learning analytics; *AIED*: Artificial intelligence in education; *LA*: Leaning analytics; *EDM*: Educational data mining; *ML*: Machine learning; *DNN*: Deep neural networks; *ANN*: Artificial neural networks; *RF*: Random forest; *GB*: Gradient boosting; *DT*: Decision tree; *RCM*: Rich context model; *MLR*: Multiple linear regression; *NB*: Naive bayes; *K-NN*: K-nearest neighbors; *SVM*: Support vector machine; *RMSE*: Root mean square error; *MCC*: Mathew correlation coefficient; *LSTM*: Long short-term memory; *MSE*: Mean squared error; *pMSE*: Predictive mean squared error; *pMAPC*: Predictive mean absolute percentage correction

Data availability The data supporting this systematic review are from previously reported studies and datasets, which have been cited. The processed studies are available at: https://github.com/NabilaSghir/systematic-review-references.git.

Declarations

Conflict of interests Not Applicable.

References

Abdullah, A. S., et al. (2021). Assessment of academic performance with the e-mental health interventions in virtual learning environment using machine learning techniques: A hybrid approach. *Journal of Engineering Education Transformations, 34*(SP ICTIEE), 79–85. https://doi.org/10.16920/jeet/2021/v34i0/157109

Adadi, A., Lahmer, M., & Nasiri, S. (2021). Artificial Intelligence and COVID-19: A systematic umbrella review and roads ahead. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.07.010

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data, 8*(1), 24. https://doi.org/10.1186/s40537-021-00419-9

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adekitan, A. I., & Salau, O. (2020). Toward an improved learning process: The relevance of ethnicity to data mining prediction of students' performance. *SN Applied Sciences*, *2*(1). https://doi.org/10.1007/s42452-019-1752-1

Adnan, M., et al. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access, 9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Afzaal, M., et al. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.723447

Albalooshi, F., AlObaidy, H., & Ghanim, A. (2019). Mining students outcomes: An empirical study. *International Journal of Computing and Digital Systems, 8*(3), 229–241. https://doi.org/10.12785/ijcds/080303

Albreiki, B., et al. (2021). Customized rule-based model to identify at-risk students and propose rational remedial actions. Big Data and Cognitive Computing, *5*(4). https://doi.org/10.3390/bdcc5040071

Almasri, A., Celebi, E., & Alkhawaldeh, R. S. (2019). EMT: Ensemble meta-based tree model for predicting student performance. *Scientific Programming*, *2019*. https://doi.org/10.1155/2019/3610248

Al-Shabandar, R., et al. (2017). Machine learning approaches to predict learning outcomes in Massive open online courses, In *Proc Int Jt Conf Neural Networks*. Institute of Electrical and Electronics Engineers Inc., pp. 713–720. https://doi.org/10.1109/IJCNN.2017.7965922

Alturki, S., Alturki, N., & Stuckenschmidt, H. (2021). Using educational data mining to predict students' academic performance for applying early interventions. *Journal of Information Technology Education: Innovations in Practice, 20*, 121–137. https://doi.org/10.28945/4835

Ayouni, S., et al. (2021). A new ML-based approach to enhance student engagement in online environment. *PLoS ONE*, *16*(11 November). https://doi.org/10.1371/journal.pone.0258788

Baek, C., & Doleck, T. (2021). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. *Interactive Learning Environments*, 1–23. https://doi.org/10.1080/10494820.2021.1943689

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics, In J. A. Larusson, & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bañeres, D., et al. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences (Switzerland)*, *10*(13). https://doi.org/10.3390/app10134427

Brooks, C. A., & Thompson, C. D. S. (2017). *Chapter 5 : Predictive Modelling in Teaching and* Learning. Available at: https://www.semanticscholar.org/paper/Chapter-5-%3A-Predictive-Modelling-in-Teaching-and-BrooksThompson/2cd4901b07f3562f98e1e56dc5712e8bc03bdc2e

Cerezo, R., et al. (2017). Procrastinating behavior in computer-based learning environments to predict performance: A case study in Moodle. *Frontiers in Psychology*, *8*(AUG). https://doi.org/10.3389/fpsyg.2017.01403

Chan, A. K., Botelho, M. G., & Lam, O. L. (2019). Use of learning analytics data in health care-related educational disciplines: Systematic review. *Journal of Medical Internet Research, 21*(2), e11241. https://doi.org/10.2196/11241

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics, 7*(2), 1–17. https://doi.org/10.18608/JLA.2020.72.1

Chen, Z., et al. (2020). Education 4.0 using artificial intelligence for students performance analysis. *Inteligencia Artificial, 23*(66), 124–137. https://doi.org/10.4114/intartif.vol23iss66pp124-137

Chui, K. T., et al. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032

Coelho, O. B., & Silveira, I. (2017). Deep learning applied to learning analytics and educational data mining: A systematic literature review. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), 28*(1), 143. https://doi.org/10.5753/cbie.sbie.2017.143

Cutad, R. E. E., & Gerardo, B. D. (2019). A prediction-based curriculum analysis using the modified artificial Bee Colony Algorithm. *International Journal of Advanced Computer Science and Applications, 10*(10), 117–123. https://doi.org/10.14569/ijacsa.2019.0101017

Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced mooc course using random forest model. *Information (Switzerland)*, *12*(11). https://doi.org/10.3390/info12110476

de Oliveira, C. F., et al. (2021). How does learning analytics contribute to prevent students dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing, 5*(4), 64. https://doi.org/10.3390/bdcc5040064

Dias, S. B., et al. (2020). DeepLMS: A deep learning predictive model for supporting online learning in the Covid-19 era. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-76740-9

Doleck, T., et al. (2020). Predictive analytics in education: A comparison of deep learning frameworks. *Education and Information Technologies, 25*(3), 1951–1963. https://doi.org/10.1007/s10639-019-10068-4

Du, X., et al. (2020). Educational data mining: A systematic review of research and emerging trends. *Information Discovery and Delivery, 48*(4), 225–236. https://doi.org/10.1108/IDD-09-2019-0070

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Ekuban, A. B., et al. (2021). *Using GitLab Interactions to Predict Student Success When Working as Part of a Team*, *Adv. Intell. Sys. Comput.* Springer Science and Business Media Deutschland GmbH, p. 138. https://doi.org/10.1007/978-3-030-68198-2_11

El Alfy, S., Marx Gómez, J., & Dani, A. (2019). Exploring the benefits and challenges of learning analytics in higher education institutions: a systematic literature review. *Information Discovery and Delivery, 47*(1), 25–34. https://doi.org/10.1108/IDD-06-2018-0018

El Aouifi, H., et al. (2021). Predicting learner's performance through video sequences viewing behavior analysis using educational data-mining. *Education and Information Technologies, 26*(5), 5799–5814. https://doi.org/10.1007/s10639-021-10512-4

Emerson, A., et al. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology, 51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Flanagan, B., Majumdar, R., & Ogata, H. (2022). Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis. *International Journal of Educational Technology in Higher Education*, *19*(1). https://doi.org/10.1186/s41239-022-00348-4

Gasevic, D., et al. (2019). How do we start? An approach to learning analytics adoption in higher education. *International Journal of Information and Learning Technology, 36*(4), 342–353. https://doi.org/10.1108/IJILT-02-2019-0024

Gitinabard, N., et al. (2019). How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Transactions on Learning Technologies, 12*(2), 184–197. https://doi.org/10.1109/TLT.2019.2911832

Goel, Y., & Goyal, R. (2020). On the effectiveness of self-training in MOOC dropout prediction. *Open Computer Science, 10*(1), 246–258. https://doi.org/10.1515/comp-2020-0153

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers and Education, 131*, 22–32. https://doi.org/10.1016/j.compedu.2018.12.006

Guo, S., Zeng, D., & Dong, S. (2020). Pedagogical data analysis via federated learning toward education 4.0. *American Journal of Education and Information Technology, 4*(2), 56. https://doi.org/10.11648/j.ajeit.20200402.13

Hasan, R., et al. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences (Switzerland)*, *10*(11). https://doi.org/10.3390/app10113894

Heilala, V., et al. (2020). Course Satisfaction in Engineering Education through the Lens of Student Agency Analytics, In *Proc. Front. Educ. Conf. FIE*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/FIE44824.2020.9274141

Herodotou, C., et al. (2019). Empowering online teachers through predictive learning analytics. *British Journal of Educational Technology, 50*(6), 3064–3079. https://doi.org/10.1111/bjet.12853

Hew, K. F., et al. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education, 145*, 103724. https://doi.org/10.1016/j.compedu.2019.103724

Hsu Wang, F. (2019). On prediction of online behaviors and achievement using self-regulated learning awareness in flipped classrooms. *International Journal of Information and Education Technology, 9*(12), 874–879. https://doi.org/10.18178/ijiet.2019.9.12.1320

Hussain, M., et al. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience*, *2018*. https://doi.org/10.1155/2018/6347186

Hussain, S., & Khan, M. Q. (2021). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*. https://doi.org/10.1007/s40745-021-00341-0

Iatrellis, O., et al. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies, 26*(1), 69–88. https://doi.org/10.1007/s10639-020-10260-x

Jensen, E., et al. (2021). What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment, In *ACM Int. Conf. Proc. Ser.* Association for Computing Machinery, pp. 121–131. https://doi.org/10.1145/3448139.3448151

Joksimović, S., et al. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning, 31*(6), 638–654. https://doi.org/10.1111/jcal.12107

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*, *11*(7). https://doi.org/10.3390/app11073130

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International*

Journal of Educational Technology in Higher Education, *18*(1). https://doi.org/10.1186/s41239-021-00300-y

Khan, I., et al. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, *8*(1). https://doi.org/10.1186/s40561-021-00161-y

Kostopoulos, G., et al. (2021). Interpretable models for early prediction of certification in MOOCs: A case study on a MOOC for smart city professionals. *IEEE Access, 9*, 165881–165891. https://doi.org/10.1109/ACCESS.2021.3134787

Kumar, A., Selvam, R., & Kumar, K. (2018). Review on prediction algorithms in educational data mining. *International Journal of Pure and Applied Mathematics, 118*, 531–536.

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology, 50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Lincke, A., et al. (2021). The performance of some machine learning approaches and a rich context model in student answer prediction. *Research and Practice in Technology Enhanced Learning*, *16*(1). https://doi.org/10.1186/s41039-021-00159-7

Liz-Domínguez, M., et al. (2019). Systematic literature review of predictive analysis tools in higher education. *Applied Sciences, 9*(24), 5569. https://doi.org/10.3390/app9245569

Lu, O. H. T., et al. (2018). Applying learning analytics for the early prediction of students academic performance in blended learning. *Journal of Educational Technology & Society, 21*(2), 220–232.

Lu, O. H. T., Huang, A. Y. Q., & Yang, S. J. H. (2021). Impact of teachers' grading policy on the identification of at-risk students in learning analytics. *Computers & Education, 163*, 104109. https://doi.org/10.1016/j.compedu.2020.104109

Macarini, L. A. B., et al. (2019). Predicting students success in blended learning-Evaluating different interactions inside learning management systems. *Applied Sciences (Switzerland)*, *9*(24). https://doi.org/10.3390/app9245523

Mai, T. T., Bezbradica, M., & Crane, M. (2022). Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. *Future Generation Computer Systems, 127*, 42–55. https://doi.org/10.1016/j.future.2021.08.026

Mansouri, T., ZareRavasan, A., & Ashrafi, A. (2021). A learning fuzzy cognitive map (LFCM) approach to predict student performance. *Journal of Information Technology Education: Research, 20*, 221–243. https://doi.org/10.28945/4760

Moher, D., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moreno-Marcos, P.M., et al. (2019). Generalizing predictive models of admission test success based on online interactions. *Sustainability (Switzerland)*, *11*(18). https://doi.org/10.3390/su11184940

Mubarak, A. A., Cao, H., & Ahmed, S. A. M. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies, 26*(1), 371–392. https://doi.org/10.1007/s10639-020-10273-6

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237. https://doi.org/10.3390/app11010237

Neha, K., Sidiq, J., & Zaman, M. (2021). Deep neural network model for identification of predictive variables and evaluation of student's academic performance. *Revue d'Intelligence Artificielle, 35*(5), 409–415. https://doi.org/10.18280/ria.350507

Nunn, S., et al. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning Journal, 20*, 1–17. https://doi.org/10.24059/olj.v20i2.790

Omer, U., Farooq, M. S., & Abid, A. (2020). Cognitive learning analytics using assessment data and concept map: A framework-based approach for sustainability of programming courses. *Sustainability (Switzerland)*, *12*(17). https://doi.org/10.3390/su12176990

Parvathi, M. (2021). Activity based analysis and prediction strategy for the class room performance improvement. *Journal of Engineering Education Transformations, 34*(Special Issue), 686–693. https://doi.org/10.16920/jeet/2021/v34i0/157167

Pedró, F., et al. (2019). Artificial intelligence in education : challenges and opportunities for sustainable development. Available at: https://www.semanticscholar.org/paper/Artificial-intelligence-in-education-%3A-challenges-Pedr%C3%B3-Subosa/697ba06bfcabbbde6292d979b87b2642115f1099

Prada, M. A., et al. (2020). Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees. *IEEE Access, 8*, 212818–212836. https://doi.org/10.1109/ACCESS.2020.3040858

Qiu, F., et al. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-021-03867-8

Rafique, A., et al. (2021). Integrating learning analytics and collaborative learning for improving student's academic performance. *IEEE Access, 9*, 167812–167826. https://doi.org/10.1109/ACCESS.2021.3135309

Raimundo, R., & Rosário, A. (2021). Blockchain system in the higher education. *European Journal of Investigation in Health, Psychology and Education, 11*(1), 276–293. https://doi.org/10.3390/ejihpe11010021

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting students performance by means of machine learning: A review. *Applied Sciences, 10*(3), 1042. https://doi.org/10.3390/app10031042

Rienties, B., Køhler Simonsen, H., & Herodotou, C. (2020). Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. *Frontiers in Education*, *5*, p. 128. Available at: https://www.frontiersin.org/articles/10.3389/feduc.2020.00128

Sghir, N., et al. (2022). Using Learning Analytics to Improve Students' Enrollments in Higher Education, in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1–5. https://doi.org/10.1109/IRASET52964.2022.9737993

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology, 50*(6), 3004–3031. https://doi.org/10.1111/bjet.12854

Shayan, P., & van Zaanen, M. (2019). Predicting student performance from their behavior in learning management systems. *International Journal of Information and Education Technology, 9*(5), 337–341. https://doi.org/10.18178/ijiet.2019.9.5.1223

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46(5), pp. 31-40. Available at: https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Spikol, D., et al. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning, 34*(4), 366–377. https://doi.org/10.1111/jcal.12263

Tuononen, T., & Parpala, A. (2021). The role of academic competences and learning processes in predicting Bachelor's and Master's thesis grades. *Studies in Educational Evaluation*, *70*. https://doi.org/10.1016/j.stueduc.2021.101001

Umer, R., et al. (2021) Current stance on predictive analytics in higher education: opportunities, challenges and future directions. *Interactive Learning Environments*, 1–26. https://doi.org/10.1080/10494820.2021.1933542

Villagrá-Arnedo, C., et al. (2016). Predicting academic performance from Behavioural and learning data. *International Journal of Design and Nature and Ecodynamics, 11*(3), 239–249. https://doi.org/10.2495/DNE-V11-N3-239-249

Waheed, H., et al. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189

Yang, S. J. H., et al. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing, 26*, 170–176. https://doi.org/10.2197/ipsjjip.26.170

Yu, C.-H., Wu, J., & Liu, A.-C. (2019). Predicting learning outcomes with MOOC clickstreams. *Education Sciences*, *9*(2). https://doi.org/10.3390/educsci9020104

Zacharis, N. Z. (2018). Classification and regression trees (CART) for predictive modeling in blended learning. *International Journal of Intelligent Systems and Applications, 10*(3), 1–9. https://doi.org/10.5815/ijisa.2018.03.01

Zawacki-Richter, O., et al. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering, 89*, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.