Education and Information Technologies
https://doi.org/10.1007/s10639-024-12704-0

A review of machine learning methods used for educational data

Zara Ersozlu[1] · **Sona Taheri**[2] · **Inge Koch**[2]

Received: 10 August 2023 / Accepted: 9 April 2024

Abstract
Integrating machine learning (ML) methods in educational research has the potential to greatly impact upon research, teaching, learning and assessment by enabling personalised learning, adaptive assessment and providing insights into student performance, progress and learning patterns. To reveal more about this notion, we investigated ML approaches used for educational data analysis in the last decade and provided recommendations for further research. Using a systematic literature review (SLR), we examined 77 publications from two large and high-impact databases for educational research using bibliometric mapping and evaluative review analysis. Our results suggest that the top five most frequently used keywords were similar in both databases. The majority of the publications (88%) utilised supervised ML approaches for predicting students' performances and finding learning patterns. These methods include decision trees, support vector machines, random forests, and logistic regression. Semi-supervised learning methods were less frequently used, but also demonstrated promising results in predicting students' performance. Finally, we discuss the implications of these results for statisticians, researchers, and policymakers in education.

Keywords Machine learning methods · Educational data analysis · Transforming educational research · Systematic review

✉ Zara Ersozlu
zara.ersozlu@newcastle.edu.au

[1] STEM Education, The University of Newcastle, Newcastle, NSW 2299, Australia

[2] Mathematical Sciences, RMIT University, Melbourne, VIC 3000, Australia

Published online: 07 May 2024

1 Introduction

Machine Learning (ML) methods have transformed the way we interact with data and have increased the potential of recognising patterns and making sense of large volumes of data. The application of ML in education is growing and has the potential to revolutionize both teaching and learning. ML methods mostly used by researchers with the purpose of predicting student's performance, analysing learning preferences and teaching effectiveness. These will not only help educators to create more effective and individualised learning opportunities for their students but also will enable educational statisticians and researchers to draw highly accurate results from educational data using ML (Hilbert et al., 2021).

ML techniques have the capacity to evaluate and learn from large volumes of data, which makes it a desirable tool for educational data. Because ML can be used to tailor instruction, evaluate learning practices, and detect patterns and trends in student learning and performance (Baker & Siemens, 2014; Kovanovic et al., 2015). The capacity of ML methods to learn from deep, non-linear correlations in data gives it advantage over traditional statistical methods (Hilbert et al., 2021). In the prediction of student performance, ML can provide better accuracy when comparing the methods of traditional statistical methods. This is because ML methods have consistently outperformed traditional methods on training data, achieving higher levels of accuracy, and generalising better across diverse datasets (Japkowicz & Shah, 2011; Kotsiantis et al., 2004).

The literature on using ML for educational data is spread across different aspects of education (e.g., students, teachers Levy et al., 2020); all schooling levels (e.g., K-12 Tedre et al., 2021); higher education (Vartiainen et al., 2022; Križanić, 2020); predicting student outcomes/performance (Khan & Ghosh, 2021; Hashim et al., 2020); learning analytics (e.g., both learning and dispositional learning analytics Buckingham Shum, & Ferguson, 2012; Gasevic et al., 2016; Tempelaar et al., 2021); early warning systems (e.g., at risk students, dropouts Pecuchová & Drlik, 2021); marking automatisation (Shermis & Burstein, 2013); language proficiency (Crossley et al., 2011) and social network analysis (Romero & Ventura, 2013).

There are also several review studies that indirectly focus on limited aspects of ML for educational data in a given timeline. Alonso-Fernández et al. (2019) have investigated game learning analytics using literature review; Bachhal et al. (2021) have discussed the most important studies conducted until 2021 in educational data mining in general; Yunita et al. (2021) has reviewed the relevant literature on big data in education; Khan and Ghosh (2021) have examined the educational data mining publications from the perspective of student performance analysis and prediction in classroom learning; Salloum et al. (2020) have analysed the literature to find out how data mining was handled by researchers in the past and the most recent trends on data mining in educational research between 2016 and 2019; Albreiki et al. (2021) have reviewed the literature on student' performance prediction using ML techniques where they focused identifying student dropouts and students at risk in literature between 2009 and 2021; Du et al. (2020) have examined 33 publications between 2007 and the first quarter of 2019 to analyse educational data mining research trends where they analysed research topics, methods and sample; Khalaf et al. (2021) have

anlaysed the literature on using only supervised ML in the period of 2010–2020; Peña-Ayala (2014) has reviewed the literature on educational data mining between 2010 and first quarter of 2013.

While the existing body of research on ML applications in educational data offers valuable insights, a closer examination reveals notable research gaps and areas where a comprehensive understanding is still elusive. In terms of the fragmentation, the majority of review studies in this area adopt a temporal scope, focusing on specific timeframes. For instance, Alonso-Fernández et al. (2019) explored game learning analytics, Bachhal et al. (2021) covered studies up to 2021, and Salloum et al. (2020) analyzed trends between 2016 and 2019. These fragmented timelines create a gap in understanding the evolution and continuity of ML applications in educational data over an extended period. From the aspect of dimensional specificity, many reviews concentrate on singular dimensions of ML applications. Yunita et al. (2021) delved into big data in education, Khan and Ghosh (2021) focused on student performance analysis, and Albreiki et al. (2021) focused on student dropouts and at-risk students. This classified approach leads to a lack of synthesis across various dimensions, leaving unexplored intersections and potential synergies. In terms of methodological variety, some reviews show limited diversity in their methodologies. While some, like Du et al. (2020), delve into specific publications, others, such as Khalaf et al. (2021), narrow their focus to supervised ML. This highlights a need for a borader approach that comprehensively synthesises the methodologies employed in existing research. Lastly, existing reviews often fall short in providing a holistic integration of ML methods tailored for educational data. While Peña-Ayala (2014) reviewed educational data mining up to the first quarter of 2013, there's a gap in synthesizing these methods comprehensively, considering advancements and changes in the landscape since then.

Addressing these research gaps is important for advancing and establishing a robust foundation for future studies in the nuanced intersection of ML applications and educational data. Our research endeavours to bridge these gaps by offering a unified, comprehensive, and contemporary analysis, thus contributing to a more holistic understanding of the subject.

Therefore, our research aims to address this limitation by conducting a thorough and comprehensive review that covers all relevant dimensions of ML methods specifically for educational data. Thus, it is imperative to provide a more comprehensive analysis of literature resources across two main databases for (Web of Science and EBSCOhost) education research using existing publications over the last decade. In this paper, we aim to investigate the existing research literature to reveal the type and range of ML approaches that have been used to analyse educational data sets. In this sense, this research is unique in terms of the aim and the practical interpretation of our findings for all educators from all schooling years and education researchers and statisticians from all backgrounds. More specifically, we aim to cover and respond to the following research questions:

1. What are the frequently used keywords and publication trends in research publications using ML to analyse educational data?

2. How can we categorise the machine learning methods utilised in research publications over the last decade, focusing on their application domains and algorithmic techniques?

2 Machine learning

ML is a branch of statistics and artificial intelligence (AI) that focuses on statistical methods to learn from data and build new statistical models and algorithms to understand, make sense of and analyse data in detail without the need for explicit programming. ML encompasses a diverse range of perspectives based on its primary applications. It can be defined as "the field of study that gives computers the ability to learn without being explicitly programmed" (Mitchell, 1997), "the process by which computers can identify patterns in data and improve their ability to recognize and predict these patterns over time" (Baker & Siemens, 2014), "a branch of artificial intelligence that systematically applies algorithms to synthesize the underlying relationships among data and information" (Awad & Khanna, 2015), and "programming computers to optimize a performance criterion using example data or past experience" (Alpaydin, 2010, p.3). Additionally, it is worth noting that ML, which was once synonymous with statistical learning until about 2015, is primarily focused on the prediction of outputs from given inputs (Hastie et al., Ch. 9, 2009; Koch, Ch. 4, 2013).

There are many other definitions of ML proposed by various researchers from different discipline backgrounds. Based on these definitions, we can infer that ML, is a branch of artificial intelligence, that employs statistical algorithms to enable computers to learn from data and iteratively improve their performance in recognising patterns and making predictions without the need for explicit programming.

It involves the systematic study and application of statistical models to analyse and synthesise the underlying patterns and connections present in data, empowering machines to make data-driven decisions and adapt to new information over time. ML utilises a range of algorithms to reveal and analyse data sets. There is no perfect algorithm that can solve every problem, each problem's complexity and nature dictate the most suitable approach for its solution. The selection of appropriate methods depends on several factors, including the problem's specific characteristics, the number of variables involved, the optimal model form and other relevant considerations (Mahesh, 2021). ML methods typically are divided into three categories:

Supervised learning In supervised learning, the algorithm is trained on labelled data in which the input characteristics are accompanied by matching output labels. The main goal is to learn a function that converts input to output (Alpaydin, 2010, Ch. 2; Hastie et al., 2009, Ch. 9). Classification (K-Nearest Neighbour (KNN), Naïve Bayes Classifier, Support Vector Machine (SVM), Logistic Regression, Linear Regression, Decision Trees, Random Forests (they are both classification and regression), Sentiment analysis etc. are mostly used in supervised learning algorithms.

Unsupervised learning In unsupervised learning, the algorithm is provided with unlabelled data, where the input features do not have matching labels on the output. Without any prior knowledge about the output, the goal is to discover patterns and structure in the data (Bishop, 2006; Goodfellow et al., 2016). Autoencoders, principal component analysis (PCA), dimension reduction and clustering (K-means), are a few examples of unsupervised learning techniques.

Semi-supervised learning Semi-supervised learning is a combination of supervised and unsupervised learning -labelled and unlabelled data (Zhu, 2008). Semi-supervised learning techniques are especially useful in situations with a shortage of labelled data to improve the reliability of results (van Engelen, & Hoos, 2020).

Reinforcement learning In reinforcement learning, an algorithm learns decision-making skills by interacting with its environment. The goal is to learn a principle that optimises a reward signal. Reinforcement learning is widely used in robotics, gaming, and control systems such as an artificial neural network (Sutton & Barto, 2018; Kaelbling et al., 1996).

ML algorithms can be further classified into several categories or variants, including deep learning. ML employs thoroughly researched and developed statistical models and algorithms to enable computer systems to iteratively refine their performance on specific tasks over time. In educational settings, these statistical approaches find application in analysing extensive and complicated data sets, revealing essential insights into students' learning patterns and preferences. As a result, ML facilitates the tailoring the learning experiences, offering personalised learning paths that align with individual requirements and consequently, optimising the educational journey for each student.

2.1 Use of ML in educational research

Studies exploring the potential uses of ML in areas such as predicting student outcomes, identifying students at risk, and customising learning experiences have all grown in recent years, with increasing interest in the use of ML in the analysis of educational data. Based on our review study, we classified the use of ML algorithms in educational research as follows:

Predicting student outcomes is one of the most exciting uses of ML methods in the examination of educational data. Researchers have employed ML methods to predict student dropout rates (Romero et al, 2008: Latham et al., 2014). Particularly, ML algorithms have demonstrated higher predictive accuracy compared to traditional approaches, highlighting the potential of ML in enhancing student performance (Hilbert et al., 2021).

Identifying at-risk students represents another area where ML has shown promise in analysing training data. As demonstrated in a study conducted by Hsu & Yeh, 2020, Rawat et al, 2021, Zhang et al, 2021 and Xing & Du 2018), ML algorithms can be used to find non-linear connections between student performance and social, as well as academic variables. The study further highlights that utilising these insights

can lead to the development of more effective interventions and activities tailored to support at-risk students' success.

Learning analytics can assists teachers to understand students' learning patterns and identify areas requiring additional support. By employing ML algorithms, trends and patterns emerge, thus enhancing teaching and learning methods and ultimately improving student learning outcomes. Within learning analytics, student behaviour and performance can be analysed to detail student's learning needs and gaps to inform teaching and assessment practices. Moreover, learning analytics encompasses dispositional learning analytics, which aims to reveal students' self-regulated learning strategies as well as their perceptions and preferences regarding specific aspects of their learning process (Buckingham Shum, & Ferguson, 2012; Gasevic et al., 2016). This comprehensive approach allows educators to gain a deeper understanding of student's individual learning journeys, leading to more effective and personalised support to foster improved learning experiences.

Natural Language Processing (NLP), as a component of ML, offers a powerful tool to analyse students' writing and extract insights about their cognitive processes, learning methods and language proficiency. Using NLP, educators can assess students' writing and provide personalised feedback to enhance their writing skills and academic achievement (Crossley et al., 2011). With the capacity to evaluate students' essays without human involvement, automated essay scoring through NLP yields trustworthy and consistent results, enabling teachers to offer targeted feedback and effectively boost students' writing abilities (Shermis & Burstein, 2013).

Adaptive learning systems are ML-based learning technologies customised to meet the specific requirements of each individual learner. By assessing student performance and dynamically adjusting the difficulty level and content of learning materials, these systems have the potential to enhance student engagement and improve learning outcomes significantly. Studies have shown that the implementation of adaptive learning systems can significantly improve students' learning outcomes (Wang et al., 2017).

Social network analysis, as an ML approach, examines the connections between individuals within a social network, offering insights into their interactions and learning patterns in the context of education. With the identification of social connections and patterns among students, this method enables educators to pinpoint student groups that may require further assistance and targeted interventions (Romero & Ventura, 2013). This powerful method provides a comprehensive understanding of the dynamics of social interactions, facilitating more effective support and guidance within educational settings.

Furthermore, ML can be effectively utilised to visualise educational data, providing insights into student performance, and learning outcomes. Statisticians can employ data visualisation techniques such as heatmaps, scatterplots, and network graphs to detect patterns and trends in student data that might not be apparent through traditional statistical analysis methods. These data visualisation tools offer researchers and statisticians the capability to uncover student activities and interactions that could significantly influence learning outcomes. By employing ML-powered visualisations, researchers and statisticians can provide educators with a deeper under-

standing of their student progress and interactions, facilitating informed decisions to enhance the overall learning experiences of students.

3 Methodology and analysis

We utilised a systematic review of research literature using both evaluative review and research mapping analysis (McBurney & Novak, 2002). An evaluative systematic review analysis was conducted using the Voytant tool, the term co-occurrence map on text data and descriptive visualisations using Excel. Web of Science and EBSCOhost (by limiting to "ERIC, Education Source, Academic Research Complete") databases were analysed for research publications (only articles) published between 2014 and 2022. We aimed to gain empirical insights by emphasizing original research findings and data-driven studies. Excluding 2023 publications is a strategic decision to ensure a focus on empirical papers where we present first-hand research results rather than synthesizing existing knowledge as seen in review articles. It was also due to the early start of our paper writing process in 2023, we have decided not to include papers published in 2023 due to their minimal presence and limited relevance to our research. This decision was made collectively in view of the need to set a cut-off point due to the timing constraints of our paper preparation.

Web of Science is recognised for its broad interdisciplinary scope, covering a wide range of academic fields. Web of Science was chosen because it is more beneficial for our research because it is more comprehensive and diverse academic publications, making it suitable for interdisciplinary studies and research that draws insights from multiple fields. Besides, citation analysis tools provided by Web of Science enable researchers a comprehensive understanding of the scientific impact of research papers. EBSCOhost encompasses various databases and indicates a strategic focus beyond education-specific literature. This decision is driven by the need for a broader exploration of topics, potentially incorporating insights from related disciplines into education.

The keywords: "machine learning", "education", "educational", "educational data", and "machine learning algorithms" were used. Using the keywords "machine learning" together or separately as "machine" and "learning" did not make a difference in terms of publication results. The same publication results were obtained in both ways.

The language of the publications was limited to only English. We included the term "machine learning" because it enabled us a broad search of publications directly related to the overarching theme of machine learning techniques in educational data. Then, we used the derivatives such as "education" and "educational" allowed for a comprehensive review of research in the field of education, ensuring that the search was not narrowly restricted to a specific aspect of education. We added the keyword "machine learning algorithms", which allowed us to go beyond general discussions to look for publications that specifically find technical aspects of machine learning in education. Once all authors had agreed on keywords and article types, the first author conducted the searches on the timeline and created the datasets.

We conducted searches on two major databases Web of Science and EBSCOhost which have high impact factors and widely read journals. After initial search we applied inclusion and exclusion criteria for each database search (as detailed in Fig. 1). The PRISMA flowchart following the suggestion of systematic review guideline by Page et al. (2021) to provide more insights into our methodological framework for this research.

On EBSCOhost, our search resulted in 320 papers initially, then we used the keyword "ML algorithms" to narrow the papers to the ones only using ML techniques specifically and this search resulted in 62 publications (48 articles, 2 conference materials and 12 reports) in total. After removing duplicates and unrelated and no empirical research articles, we only focused on the scientific articles in this study, therefore 27 research articles in total were included in the analysis from EBSCOhost databases in total.

On Web of Science, the first search yielded 560 research articles initially. We limited the articles to the area of education and education research 77 articles were found. After using the keyword "ML algorithms" our search resulted in 62 articles, we checked if all papers were related to education, and if they used empirical research articles. We removed the irrelevant papers, non-empirical research and the duplicates removed those from the file which reduced the number of articles to 50. Consequently, we ended up 77 as a collective result from both databases. We used the Voyant tool for mapping the keywords whereas we utilised excel for evaluative review analysis. We classified and coded the papers based on their application domains, methodologies and statistical techniques used.

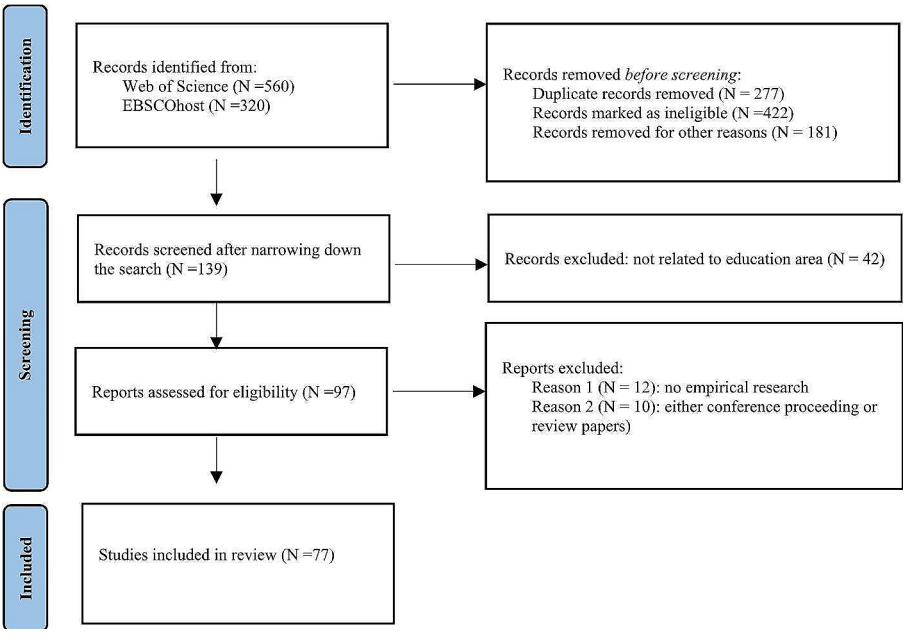

Fig. 1 PRISMA flowchart for this study's methodology

4 Results

In this part, each research question is responded to by using evaluative analysis of the literature review of ML in the education field. The results below are specified based on the research questions.

4.1 What are the frequently used keywords and publication trends in research publications using ML to analyse educational data?

To address the first question, we mapped the keywords analysis using Voyant tool and descriptive visualisations using Excel to summarise the range of the articles based on their year of publication and publication databases to reveal publication trends. The word "learning" was the most frequent word in both corpuses, indicating that the word learning is quite commonly used for ML analytics in both databases (Fig. 2). Similarly, the word "machine" was among the top five most frequently used words in both databases, indicating that ML is popular for educational data mining in both databases.

The words "students", "predicting" and "performance" followed the words "learning" and "machine" in the top five most frequent words. This could suggest that research publications in both databases have focused on predicting student performance of students.

Fig. 2 Most frequent words in the corpus combined from two databases

[Most frequent words were in the corpus for Web of Science database: learning (67); machine (37); students (27); predicting (25); performance (22).

Based on Fig. 3, between the year range from 2014 to 2022, majority of the papers were published between 2021 and 2022 in both databases. The increase in publications starting from 2019 can suggest a growing interest to using ML algorithms in education research. In both databases, the concentration of publications between 2018 and 2022 is noteworthy and indicates that the number of publications using ML analytics is developing rapidly (Fig. 3). Furthermore, the years 2020 and 2021 seem to be particularly popular for publications, with a high number of publications are in these years.

The earliest publication date was 2014. This suggests that the interest in using ML for educational data may be relatively new. It can also indicate that both databases may only have started collecting publications on this topic in 2014.

4.2 How can we categorise the machine learning methods utilised in research publications over the last decade, focusing on their application domains and algorithmic techniques?

Based on our analysis of papers, we created two associated themes based on used ML algorithms and the application domains using these algorithms. Excel was used to analyse and visualise the data and results. As we analysed articles around the themes we determined, we used triangulation technique to compare ML algorithms to interpret most common themes for both application domains and specific ML algorithms used for educational data analytics. To ensure the coding reliability, another data scientist independently coded 77 articles based on the application domains of ML analysis and the actual ML analysis used. 94.9% agreement established after comparing two coding schemes which exceeds the suggested 80% reliability criteria by Miles et al. (2014). Based on the agreed description for codes below, we analysed the data further using excel visualisation techniques such as charts, graphs, and pivot tables (Tables 1 and 2).

We used a coding style that described type of the ML methods. For example: SMLA (Supervised ML algorithm), SSMLA (semi-supervised ML algorithm), RMLA (Reinforcement ML algorithm) and USMLA (Unsupervised ML algorithm). We combined our analysis from both databases to see the most frequently used ML

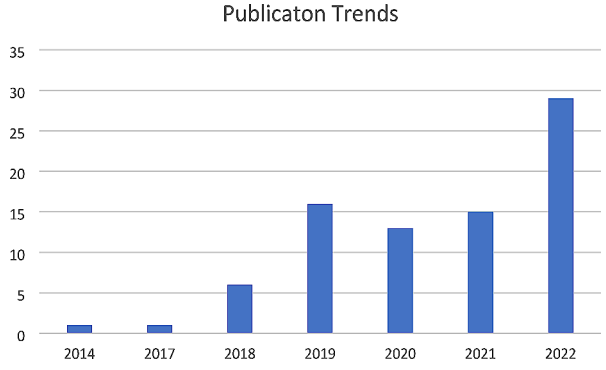

Fig. 3 Combined publication trends from both databases

Table 1 Application domains

Application domains descriptions:	
PT1	Predicting student performance
PT2	Prediction of student drop out
PT3	Learning analytics
PT4	Teaching quality assessment

Table 2 Type of ML methods

ML methods descriptions:	
SMLA1	Linear regression
SMLA2	Gradient boosting
SMLA3	Random forest
RMLA1	Neural networks, deep learning
SMLA4	Sentiment mining, natural language processing
SMLA5	Support vector machine classifier
SMLA6	Decision tree
SMLA7	Logistic regression
SMLA8	K-Nearest neighbours
SMLA9	Naïve Bayes
SSMLA1	Semi-supervised learning
SMLA10	NNge (classification)
SMLA11	Quadratic discriminant analysis
SMLA12	Multikernel learning
SMLA13	Feature selection
USMLA2	Principal component analysis
USMLA1	Unsupervised learning

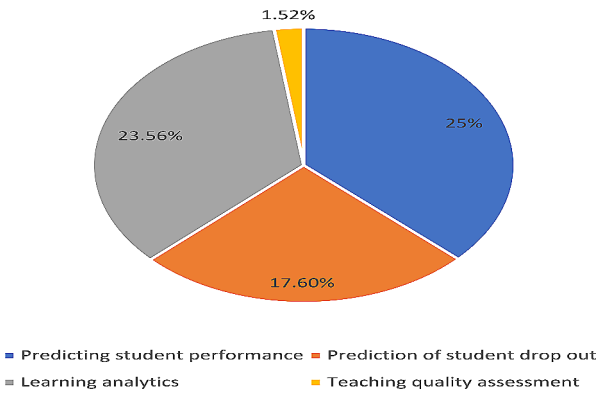

Fig. 4 Application domains used in ML research in education

algorithms and their application domain for educational data. Below pie charts visualise these results. 25% of the combined publications were aiming to predict student performance (PT1), 17.6% of them were aiming to predict student dropouts while 23.56% of the publications was focusing on learning analytics and finally 1.52% of them were targeted to measure teaching quality (Fig. 4).

In terms of the type of ML used in publications (Fig. 5), large proportionate of publications (88% in total) used supervised learning algorithms (linear regression:

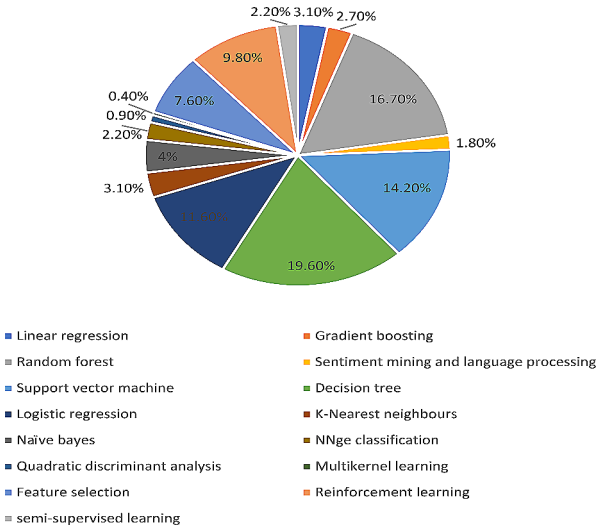

Fig. 5 Type of ML Methods used in ML research in education

Table 3 Results from both databases for the application domains of ML methods in educational research

Application domains	Articles from both databases (numbered)
Predicting student performance (PT1)	*2014*: [28]; 2017: [1]; *2018*: [2], *2019*: [4], [5], [6], [8], [32], [35], [38], [39]; *2020*: [10], [12], [13], [15], [16], [40], [42]; *2021*: [19], [45], [46]; *2022*: [24], [26], [56], [57], [59], [61], [63], [66], [67], [72], [73].
Prediction of student drop out (PT2)	*2018*: [30]; *2019*: [31], [36], [43], [51], [52], [58], [64], [71]; *2021*: [17].
Learning analytics (PT3)	*2018*: [3]; *2019*: [7], [33], [34], [37], [44], [47], [49], [50], [53], [54], [55], [60], [62], [65], [68], [69], [70], [74], [75], [76], [77]; *2020*: [11], [14]; *2021*: [18], [20], [21]; *2022*: [22], [23], [25].
Teaching quality assessment (PT4)	*2020*: [41]; *2022*: [27].

3.1%; gradient boosting: 2.7%; random forest: 16.7%; Sentiment mining and language processing: 1.8%; Support vector machine: 14.2%; Decision tree: 19.6%; Logistic regression: 11.6%; K-nearest neighbours: 3.1%; Naïve Bayes: 4%; NNge classification: 2.2.%; quadratic discriminant analysis: 0.9%; multikernel learning: 0.4%; feature selection: 7.6%).

Reinforcement learning was employed in 9.8% of the remaining 12% of the papers through neural network analytic approaches. Only 2.2% of the publications used unsupervised or semi-supervised learning algorithms through principal component analysis and some combinations of supervised unsupervised learning algorithms. The pivot tables we created that mapped out the ML methods per publications across years of publications (see the supplementary document for more detail).

We presented the results of both databases on these tables to provide clearer picture of dispersal of publications across ML methods and their main domain of use.

Table 4 Results from both databases for ML methods used in education research

ML algorithm with codes	Articles from both databases (numbered)
Linear regression (SMLA1)	*2014*: [28]; *2019*: [35]; *2021*: [21], [46], [47], *2022*: [56], [61].
Gradient boosting (SMLA2)	*2020*: [43]; *2021*: [46]; 2022: [63], [76]
Random forest (SMLA3)	*2017*: [1]; 2018: [2], [29]; *2019*: [32], [33], [39]; 2020: [9], [10], [11], [13], [16], [43], [44]; *2021*: [21], [46], [47], [49], [50], [52], [53], [54]; *2022*: [22], [24], [58], [59], [60], [62], [63], [64], [65], [66], [68], [72], [74], [75], [76], [77]
Neural networks, deep learning (RMLA1)	*2018*: [2]; *2019*: [8], [36], [37]; *2020*: [42]; *2021*: [20], [21], [45], [47], [52], [54]; *2022*: [23], [24], [57], [58], [59], [67]
Sentiment mining (SM), Natural language processing (SMLA4)	*2014*: [28]; [33], [34]; *2019*: [7]; *2022*: [55], [70]
Support vector machine (SMLA5)	*2018*: [2]; *2019*: [32], [33], [37], [39]; *2020*: [10], [12], [13], [14], [40], [41], [44]; *2021*: [20], [45], [49], [50], [51], [52], [53]; *2022*: [22], [23], [24], [26], [27], [57], [60], [61], [62], [66], [72], [73], [77]
Decision tree (SMLA6)	*2014*: [28]; 2018: [2], [3], [29]; *2019*: [4], [8], [32], [33], [35], [39]; *2020*: [11], [13], [40], [43], [44]; *2021*: [20], [21], [46], [47], [49], [50], [51], [52], [53], [54]; *2022*: [24], [57], [60], [61], [62], [66], [72], [73], [77]
Logistic regression (SMLA7)	*2018*: [29], [30]; 2019: [37]; *2020*: [10], [11], [13], [43]; *2021*: [19], [21], [51]; *2022*: [22], [24], [57], [58], [59], [60], [61], [62], [64], [67], [68], [69], [71], [72], [75], [77]
K-Nearest neighbours (SMLA8)	*2019*: [32], [39]; *2020*: [10], [13], [14], [40]; *2021*: [45]; *2022*: [24], [26], [73]
Naïve bayes (SMLA9)	*2019*: [33]; *2020*: [13], [14], [44]; *2021*: [20], [50], [51]; *2022*: [24]
Semi-supervised Learning (SSMLA1)	*2019*: [5], [6], [32], [38]; *2021*: [48]
NNge classification (SMLA10)	
Quadratic discriminant Analysis (SMLA11)	*2019*: [10]; *2020*: [21]; *2022*: [25]
Multikernel learning (SMLA12)	*2022*: [27]
Feature selection (SMLA13)	*2014*: [28]; *2018*: [30]; *2019*: [31], [34], [37], [38], [39]; *2020*: [10], [16]; *2021*: [45], [48]; *2022*: [56], [62], [64], [67], [74]
Principal component analysis (USMLA2)	*2020*: [16]
Unsupervised learning (USMLA1)	*2022*: [25]

Tables 3 and 4 show the results of our review of research studies investigating the use of ML methods to predict various aspects of student academic performance and learning behaviours, including academic grades and performance level, student drop-out, teaching quality assessment and learning analytics. The studies were conducted between 2014 and 2022 and represent a range of educational levels, from secondary education to higher education.

These tables indicated that the number of studies using ML methods has increased steadily over the past few years, with 1 study published in 2014, and a total of 76 studies published between 2017 and 2022. The results of the studies suggested that ML methods can be effective tools for predicting student academic performance, with 32 studies reporting success in predicting academic grades, and 30 studies explored their use in learning analytics. In addition, the studies highlighted the potential of

these methods for use in teaching quality assessment and predicting student drop out. However, there is still much work to be done in this area, as only one study in the table explored the use of these methods for teaching quality assessment (2 studies), and only 10 studies reporting predicting student drop out.

From Tables 3 and 4, it can be observed that several ML methods have been employed in educational data mining studies. The most used methods were random forest (37 studies), decision tree (34 studies), logistic regression (26 studies) and support vector machine classifier (32 studies). These methods have been used in several studies, indicating their effectiveness in predicting academic performance, and student drop out. SVM, decision tree and random forest were heavily used for predicting student performance while logistic regression is mostly used for learning analytics. Neural networks and deep learning (17 studies) were largely used for learning analytics and predicting student performance. Feature selection (16 Studies) was mostly used for learning analytics following predicting student performance. KNN (10 studies) was mostly used to predict student performance and Naïve Bayes (8 studies) was used in balance across predicting students' performance, learning analytics and student drop out. Additionally, the use of sentiment mining (SM) language processing in educational data mining was also notable. It was heavily used for learning analytics based on our results. The approach has been employed in several studies, and its effectiveness in learning analytics was impressive.

Table 4 shows that several studies employed multiple methods, indicating that combining methods can improve the accuracy of predictions. It was further evident that unsupervised learning approaches were not popular in educational data mining, as the table shows that only two studies used this approach. The findings indicated that there was a growing interest in using ML techniques to predict students' academic success and analytics or learning. The most used methods in these studies were random forest, support vector machine, decision tree, and neural networks. These methods were found to perform well in predicting student performance and finding patterns in learning in most of the studies.

Semi-supervised learning and unsupervised learning were used less frequently but still showed promise in predicting student performance in some studies. Additionally, sentiment mining was found to be a useful approach for analysing students' attitudes and behaviours in collaborative learning environments. It is worth noting that some studies utilised multiple methods in their analyses, which highlights the importance of selecting the appropriate methods for specific educational contexts and research questions. The use of ML methods and data mining techniques in educational research can enable educators to gain insights into student learning patterns and develop personalised interventions that can improve student outcomes.

5 Discussion

Interest in using ML for educational data has grown significantly over the last decade. According to our research, ML methods have been more frequently used with the purpose of learning analytics and prediction of student performance more frequently in education research to guide educators' decision-making, which has an impact on

all stakeholders. According to Long and Siemens (2011), there is a growing trend towards the use of data analytics and predictive modelling in education to promote student performance and improve educational outcomes.

Although the word "learning" is the most frequent word in both databases, publications in both databases emphasise on learning analytics and predicting student performance because the word "students", "predicting", "performance" are used more frequently in both databases. On the other hand, "machine" and "learning" are among the top five most frequently used words in both databases, which may indicate a focus on data mining and a broader range of scientific research topics.

A large proportion of papers in both databases were published between 2019 and 2022, demonstrating an ongoing interest in employing ML methods in educational research. The number of publications employing ML methods is growing quickly in the EBSCOhost database, with a focus on articles published between 2019 and 2022. There is a growing interest from researchers to utilise ML methods for educational data since 2019.

The results of our research show that ML methods are more frequently used in educational data mining to predict various elements of students' academic performance and learning habits. Most of the articles (88%) used supervised learning methods, which are the most frequently used methods. Decision tree, support vector machine classifier, random forest and logistic regression were the most commonly used supervised learning methods. This result was comparable to that of Luan and Tsai (2021), whose research revealed that the top 50 studies on AI in higher education. They found that these studies mainly employed traditional ML techniques such as linear regression, support vector machines, classification and clustering, data mining. Similarly, Issah et al. (2023) found that classification and decision trees are the most widely used methods in predicting student performance.

The supervised learning methods we explored are found to be successful in predicting student academic grades and student dropout which parallels a research study result found by Khalaf et al. (2021). Some other research furthermore suggests that these methods have been found to be more effective in predicting student outcomes in a range of contexts than traditional classroom settings (Qiu et al., 2021).

There is also evidence to suggest that semi-supervised learning methods, which combine labelled and unlabelled data, can be particularly effective in predicting student outcomes when labelled data is limited (Livieris et al., 2019). Additionally, some studies have found that feature selection techniques can improve the performance of ML models in predicting student outcomes (Xiao et al., 2021). However, it is worth noting that the effectiveness of ML methods in predicting student outcomes can be influenced by several factors, including the quality and quantity of data available, the context of the study, and the specific method and model used (Zaffar et al., 2018).

Only 9.8% of the publications employed reinforcement learning via neural network analysis algorithm. With this algorithm, models are trained to make choices based on incentives or punishments. To create personalised adaptive learning systems that cater to the needs of specific learners, reinforcement learning algorithm can be very helpful.

Romero and Ventura (2010) have suggested that there is a need for traditional mining algorithms to be adjusted to accommodate for the context of education. Because

data mining algorithms must take semantic information into account when analysing educational data. They have suggested that this highlights the need for more efficient mining tools that include educational field expertise into data mining algorithms. Based on our results, 1.8% of publications utilised sentiment mining and language processing. Particularly analysing verbal or written language to find and interpret attitudes, opinions, and feelings of students. This method was widely utilised for learning analytics and investigating the attitudes and behaviours of students in group learning settings (Chen et al., 2020).

A study by Japkowicz & Shah (2011) has compared the effectiveness of traditional statistical methods with ML methods for predicting student performance. The results of the study showed that ML methods continuously outperformed conventional methods, reaching greater levels of accuracy and better generalisation across various datasets. One advantage of ML methods over classical statistical techniques is their ability to learn from complex, non-linear relationships in data. Another study by Kotsiantis et al. (2004) have compared the performance of six ML methods including SVM, logistic regression, 3NN, SMO, Naïve Bayes etc. in predicting student performance. The study found that the Naive Bayes method exhibits highly satisfactory accuracy when compared to other algorithms, it stands out as the simplest one to implement. However, it is essential to exercise caution while using the Naive Bayes algorithm since its appropriateness may vary depending on the specific characteristics of the data and the nature of the problem. Proper consideration and understanding of the data and problem context are crucial when deciding whether to employ the Naive Bayes algorithm in specific applications.

Based on our findings, we determined that predictive modelling is largely used with educational data. Predictive models are not a new statistical technique to education statisticians and researchers. Even though the challenge of learning prediction models from data is the same for both supervised ML and inference statistics, and they are both based on the same mathematical ideas, supervised ML focuses on predictive modelling via non-parametric models (Hilbert et al., 2021). The main question is why ML should be used for analysing educational data analytics? Among many other contributions, all research studies we examined supported and suggested the notion of using ML methods over classical test theory techniques because of their power of accuracy and detecting stronger predictors to generalise beyond the sample and the fairness it brings to statistical analysis versus classical test theory which mainly focuses on finding correlations among variables that most of the time remain short in terms of accuracy of predictions. There are of course challenges using ML methods for educational data mostly raised by interpretability of ML methods for educational results. There are ethical and algorithmic challenges when balancing human- and machine-assisted learning (Luan et al., 2020). One notable challenge is the need for comprehensive and high-quality data to effectively train models (Mitchell, 1997). Training datasets can be complex, heterogeneous and lack standardization, making it difficult to derive meaningful insights (Lindl et al., 2020; Rudin et al., 2022). Furthermore, interpretability of machine learning models in the educational context is crucial, as stakeholders, including educators, managers and researchers need to understand the decision-making processes of these algorithms (Hilbert et al., 2021). Ensuring algorithmic fairness and reducing bias is another major hurdle,

as models may unintentionally perpetuate or even exacerbate existing inequalities in the education system. Hence ethical concerns about the privacy and security of student data require careful consideration and robust safeguards. Integrating machine learning into educational practice requires collaboration between data researchers, educators, managers and policy makers to overcome these challenges and harness the full potential of machine learning while ensuring responsible and equitable use for educational data.

Furthermore, based on research studies we examined, ML methods play an important role in predicting student performance, detecting patterns in student's learning, attitudes and dispositions as well as predicting students at risk and dropout rates (Albreiki et al., 2021). ML methods can enhance the overall quality of data analysis in educational research and demonstrate how ML can play a significant role in the validation of empirical models (Hilbert et al., 2021).

6 Conclusion, implications and limitations

Our comprehensive analysis has provided a snapshot of the current state of ML methods in the field of educational data. The databases that we selected contain widely read and cited journals covering ML methods in educational data to provide a comprehensive view of the mainstream research perspective on the application of ML methods in educational contexts. By including previous review studies in our analysis, we aimed to deepen our understanding and provide a more nuanced interpretation of our findings.

The collective findings from the analysed research publications strongly suggest that machine learning methods have demonstrated remarkable effectiveness in predicting student performance, identifying patterns and learning needs, and identifying at-risk students. The implications of these findings for educators are profound, as the availability of such knowledge can significantly transform teaching, learning and assessment practices. Personalised and adaptive approaches to education are emerging and moving away from the traditional one-size-fits-all paradigm. One way to do this is to increase educational statisticians and researchers' awareness and knowledge of ML methods to further their data analysis, as these results influence the decision-making process for all stakeholders.

However, it is important to acknowledge some limitations in our study. While our analysis sheds light on the current landscape, it is not comprehensive and may not capture new trends or the latest developments in the rapidly evolving field of machine learning in education. Not including 2023 publications in our study may create a temporal bias and future research should consider including more recent publications to provide a comprehensive understanding. Furthermore, the successful application of machine learning methods in education depends on the awareness and expertise of educational statisticians and researchers. There is a need to bridge the gap between traditional educational research and advanced data analysis techniques. Future research efforts should explore strategies that will enhance the knowledge and skills of educational stakeholders and empower them to use machine learning methods effectively. This interdisciplinary collaboration between educators and data

scientists has the potential to optimise decision-making processes for all stakeholders involved in the education ecosystem.

The accuracy of predictions based on educational data is crucial, as the results of such analyses can drive education policies worldwide, especially for international exams. In analysing well-known international tests such as PISA, PIRLS, TIMMS and TALIS, ML methods can be used to improve the accuracy of predictions and reduce biases that naturally arise from the data. These publications we analysed in this current research help to improve the interpretability of ML methods in educational research. We recommend that statisticians, researchers, educators, and policymakers collaborate to develop guidelines and policies for ethical and responsible use of ML methods in education.

In conclusion, our study highlights the transformative potential of machine learning methods in reshaping education and calls for a concerted effort to bridge the gap between classical statistical test theory techniques for educational data and the latest data analysis techniques such as ML techniques. As we navigate the evolving landscape of machine learning in education, continued research and collaboration will be instrumental in realising the full potential of using ML techniques for educational data.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Zara Ersozlu. The first draft of the manuscript was written by Zara Ersozlu, Sona Taheri and Inge Koch commented on all versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability Please note that all data generated or analysed during this study are included in this article as an appendix [and its supplementary information files].

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.
All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
The authors have no financial or proprietary interests in any material discussed in this article.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552. https://doi.org/10.3390/educsci11090552.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612.

Alpaydin, E. (2010). *Introduction to machine learning*. The MIT.

Awad, M., & Khanna, R. (2015). *Efficient learning machines theories, concepts, and applications for engineers and system designers*. A.

Bachhal, P., Ahuja, S., & Gargrish, S. (2021). Educational data mining: A review. *Journal of Physics: Conference Series*, *1950*(1), 012022. https://doi.org/10.1088/1742-6596/1950/1/012022.

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. https://doi.org/10.1017/CBO9781139519526.016.

Bishop, C. M. (2016). *Pattern recognition and machine learning*. Springer.

Buckingham Shum, S., & Ferguson, R. (2012). Social Learning Analytics. *Journal of Educational Technology & Society*, *15*(3), 3–26. http://www.jstor.org/stable/jeductechsoci.15.3.3.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002.

Crossley, S. A., Allen, D., & McNamara, D. S. (2011). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*(1), 89–108. https://doi.org/10.1177/1362168811423456.

Du, X., Yang, J., Hung, J. L., & Shelton, B. (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.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.

Hashim, A., Akeel, W., & Khalaf, A. (2020). Student performance prediction model based on supervised machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *928*(3), 032019. https://doi.org/10.1088/1757-899X/928/3/032019.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of statistical learning: Data mining, inference, and prediction* (2nd ed., Vol. 2, pp. 1–758). Springer.

Hilbert, S., Coors, S., Kraus, E. B., Bischl, B., Frei, M., Lindl, A., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). *Machine Learning for the Educational Sciences*. https://doi.org/10.31234/osf.io/3hnr6.

Hsu, C. & Yeh, C. (2020). "Mining the student dropout in higher education." ASTM International. *Journal of Testing and Evaluation. 48*(6), 4563–4575. https://doi.org/10.1520/JTE20180021.

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*, *7*, 100204. https://doi.org/10.1016/j.dajour.2023.100204.

Japkowicz, Nathalie & Shah, Mohak. (2011). Evaluating learning algorithms: A classification perspective. Evaluating Learning Algorithms. A Classification Perspective. https://doi.org/10.1017/CBO9780511921803.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Khalaf, A., Dahr, J. M., Najm, I. A., Kamel, M. B. M., Hashim, A. S., Akeel, W. A., & Humadi, M. A. (2021). Supervised learning algorithms in educational data mining: A systematic review. *Southeast Europe Journal of Soft Computing*, *10*, 55–70. https://doi.org/10.21533/scjournal.v10i1.199.

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, *26*(1), 205–240. https://doi.org/10.1007/s10639-018-9784-6.

Koch, I. (2013). *Analysis of multivariate and high-dimensional data*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025805.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*, 411–426.

Kovanovic, V., Gasevic, D., Joksimovic, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *Internet and Higher Education*, *27*, 74–89. https://doi.org/10.1016/j.iheduc.2015.06.002.

Križanić, S. (2020). Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*, 12. https://doi.org/10.1177/1847979020908675.

Latham, A., Crockett, K., & McLean, D. (2014). An adaptation algorithm for an intelligent natural language tutoring system. *Computers & Education*, *71*, 97–110. https://doi.org/10.1016/j.compedu.2013.09.014.

Levy, J., Mussack, D., Brunner, M., Keller, U., Cardoso-Leite, P., & Fischbach, A. (2020). Contrasting classical and machine learning approaches in the estimation of value-added scores in large-scale educational data. *Frontiers in Psychology*, *11*, 2190. https://doi.org/10.3389/fpsyg.2020.02190.

Lindl, A., Krauss, S., Schilcher, A., & Hilbert, S. (2020). Statistical methods in transdisciplinary educational research. *Frontiers in Education*, *5*, 97. https://doi.org/10.3389/feduc.2020.00097.

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research*, *57*(2), 448–470. https://doi.org/10.1177/0735633117752614.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30–41.

Luan, H., & Tsai, C. C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, *24*(1), 250–266. https://www.jstor.org/stable/26977871.

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C. C. (2020). Challenges and future directions of Big Data and Artificial Intelligence in Education. *Frontiers in Psychology*, *11*, 580820. https://doi.org/10.3389/fpsyg.2020.580820.

Mahesh, B. (2021). Machine learning algorithms - a review. *International Journal of Engineering and Advanced Technology*, *10*(6), 2109–2113. https://doi.org/10.35940/ijeat.F1543.1196S621.

McBurney, M., & Novak, P. (2002). What is Bibliometrics and why should you care? *IEEE International Professional Communication Conference*, 108–114. https://doi.org/10.1109/IPCC.2002.1049094.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: A methods sourcebook (3rd ed.). London, UK: SAGE.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. Revista espanola de cardiologia (English ed.), 74(9), 790–799. https://doi.org/10.1016/j.rec.2021.07.010.

Pecuchová, J., & Drlík, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*, 3130. https://doi.org/10.3390/app11073130.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.062.

Qiu, F., Zhang, G., Sheng, X. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports* 12: 453. https://doi.org/10.1038/s41598-021-03867-8.

Rawat, S., Kumar, D., Kumar, P., Khattri, C. (2021). A systematic analysis using classification machine learning algorithms to understand why learners drop out of MOOCs. *Neural Computing and Applications, 33*, 14823–14835. https://doi.org/10.1007/s00521-021-06122-3.

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

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

Romero, C. & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowl Discov, 3*, 12–27. https://doi.org/10.1002/widm.1075.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys Statist Surv*, *16*, 1–85.

Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020). Mining in Educational Data: Review and Future Directions. In: Hassanien, AE., Azar, A., Gaber, T., Oliva, D., Tolba, F. (Eds.) Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). AICV 2020. *Advances in Intelligent Systems and Computing*, vol 1153. Springer, Cham. https://doi.org/10.1007/978-3-030-44289-7_9.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (1st ed.). Routledge. https://doi.org/10.4324/9780203122761.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.).). MIT Press.

Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching machine learning in K-12 Computing Education: Potential and pitfalls: Pedagogical and Technological trajectories for Artificial Intelligence Education. *Ieee Access: Practical Innovations, Open Solutions*, 1–1. https://doi.org/10.1109/ACCESS.2021.3097962.

Tempelaar, D., Rienties, B., & Nguyen, Q. (2021). The contribution of dispositional learning analytics to precision education. *Educational Technology & Society*, *24*(1), 109–122. https://www.jstor.org/stable/26977861.

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*, 373–440. https://doi.org/10.1007/s10994-019-05855-6.

Vartiainen, H., Pellas, L., Kahila, J., Valtonen, T., & Tedre, M. (2022). Pre-service teachers' insights on data agency. *New Media & Society Advance Online Publication*. https://doi.org/10.1177/14614448221079626.

Wang, Y., Liu, X., & Chen, Y. (2017). Analyzing cross-college course enrollments via contextual graph mining. *PloS one, 12*(11), e0188577. https://doi.org/10.1371/journal.pone.0188577.

Xiao, W., Ji, P., & Hu, J. (2021). RnkHEU: A hybrid feature selection method for predicting students' performance. *Scientific Programming*. https://doi.org/10.1155/2021/1670593.

Xing, W., & Du, D. (2018). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research. 57*. 073563311875701. https://doi.org/10.1177/0735633118757015.

Yunita, A., Santoso, H. B., & Hasibuan, Z. A. (2021). Research review on big data usage for learning analytics and educational data mining: A way forward to develop an intelligent automation system. *Journal of Physics: Conference Series*, *1898*(1), [012044]. https://doi.org/10.1088/1742-6596/1898/1/012044.

Zhang, J., Gao, M., & Zhang, J. (2021). The learning behaviours of dropouts in MOOCs: A collective attention network perspective. *Computers & Education, 167*, Article 104189. https://doi.org/10.1016/j.compedu.2021.104189.

Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. (2018). A study of feature selection algorithms for predicting students' academic performance. *International Journal of Advanced Computer Science and Applications*, *9*.

Zhu, X. (2008). Semi-supervised learning literature survey. Technical Report. 1530, University of Wisconsin Madison.