Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 24/11/2025; 1st round notif.: 04/02/2026; New version: 23/02/2026; 2nd round notif.: 25/02/2026;
Camera ready: 11/03/2026; Edition review: 25/03/2026; Available online: 25/03/2026; Published: 25/03/2026;

Model Interpretability in the Educational Data Mining Context: A Systematic Literature Mapping

Cássio Soares Carvalho
Graduate Program in Computing
Federal University of Pelotas
ORCID: 0009-0003-0207-9023
cassio.carvalho@inf.ufpel.edu.br

Júlio Carlos Balzano de Mattos
Graduate Program in Computing
Federal University of Pelotas
ORCID: 0000-0002-0619-9271
julius@inf.ufpel.edu.br

Marilton Sanchotene de Aguiar
Graduate Program in Computing
Federal University of Pelotas
ORCID: 0000-0002-5247-6022
marilton@inf.ufpel.edu.br

Abstract

Machine learning (ML) techniques in the Educational Data Mining (EDM) context enable the development of increasingly efficient prediction models. This evolution benefits from the improvement of ML techniques, the current massive collection of heterogeneous data, and the high availability of computational power. As a result, efficient models are obtained; however, they are complex and challenging to comprehend. Therefore, the field of model interpretability, also known as Explainable Artificial Intelligence (XAI) research, becomes a pivotal piece in consolidating and giving credibility to the solutions developed in the EDM context. In this sense, we presented a systematic mapping to understand how studies in the area of EDM address interpretability, aiming to answer questions related to the (i) interpretability context, (ii) interpretability methods, metrics, and objectives, (iii) education levels, (iv) educational data, and (v) ML techniques. To achieve this goal, we conducted a Systematic Literature Mapping (SLM), which involved defining a research protocol with planning, conducting, and reporting phases. These phases included defining research questions, establishing digital reference libraries, and establishing inclusion and exclusion criteria. The findings indicate that, despite the peculiarities of each study, interpretability is frequently addressed as a post hoc component rather than as a core objective of model design, limiting its systematic evaluation and comparative analysis across models. There is a need for studies in which interpretability is a central objective, including the comparison of interpretability across models from different ML techniques, the exploration or proposal of interpretability metrics (particularly agnostic ones), and the investigation of the relationship between interpretability and algorithmic fairness. Overall, this study offers a comprehensive perspective on the applicability of interpretability methods in various educational contexts, synthesizes best practices and limitations in measuring and comparing model interpretability, and highlights the importance of involving stakeholders in the development of transparent and effective EDM applications.

Keywords: *Educational Data Mining; Interpretability; Explainability; Explainable AI; Fairness; Systematic Literature Mapping; Systematic Literature Review*

Cite as: Carvalho, C. S., Mattos, J. C. B., & Aguiar, M. S. (2026). Model Interpretability in the Educational Data Mining Context: A Systematic Literature Mapping. Revista Brasileira de Informática na Educação, vol. 34, pp. 314–355. https://doi.org/10.5753/rbie.2026.6984.

1 Introduction

Artificial intelligence (AI) has already become ubiquitous and is responsible for many decisions in today's society. Due to the wide availability of heterogeneous data and computational power, Machine Learning (ML) algorithms achieve increasingly better predictive performances. However, most of the models generated have greater complexity and limitations such as opacity or lack of transparency, which inherently characterizes black box models (D. V. Carvalho et al., 2019; Linardatos et al., 2021).

Once the internal logic and inner workings of these black box models are hidden, humans cannot interpret or understand the system's reasoning and how decisions are made (Montavon et al., 2017). Systems with those characteristics are complex to be trusted, especially in areas where moral and fairness issues are involved (Linardatos et al., 2021). Regarding high-stakes decisions, the stated problem is further compounded because entrusting essential decisions to a system that cannot explain itself and cannot be explained by humans presents evident dangers (Adadi & Berrada, 2018).

Research on explainable models is required in various fields, including transportation, healthcare, legal (criminal justice), military, cybersecurity, education, entertainment, government, and image recognition, among others (Adadi & Berrada, 2018). Specifically, in the case of education, several pitfalls arise from using black box models, including a lack of transparency, potential for bias, limited interpretability, dependence on data quality, and difficulty adapting to changing circumstances (Samek et al., 2019). To address those issues, Explainable Artificial Intelligence (XAI) emerged as a field of study to create a suite of interpretable models and methods that produce more explainable models while preserving high predictive performance levels (Adadi & Berrada, 2018). Yet, interpretability in the ML context aims to prevent society from being harmed or marginalized by the current generation of AI systems (Vieira & Digiampietri, 2022).

Interpretability and explainability are closely related concepts. Some authors usually differentiate them. When humans can understand what a model has done, it is interpretable. When considering that humans can explain and discover why certain attributes have a significant influence on the result, then the model is said to be explainable (Burkart & Huber, 2021). In the context of this work, as well as in Molnar (2022), the terms interpretability and explainability will be used interchangeably.

Different criteria can be used to classify methods and techniques for ML Interpretability. Intrinsic versus post hoc is another example. These criteria distinguish whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post hoc). Short decision trees (DT) or sparse linear models, for example, are considered interpretable due to their simple structure. Post hoc interpretability refers to the application of interpretability methods after a model has been trained and evaluated. Permutation feature importance and the Global Surrogate Model are examples of post hoc interpretability methods (Molnar, 2022).

Another criterion is model-specific versus model-agnostic. These classification criteria consider which class of models the interpretability methods can be applied to. If an interpretability method is based on the internal structure of a specific model or a class of models, then that method is model-specific. On the other hand, if the interpretability method can be applied to any class of

models, then it is classified as model-agnostic (D. V. Carvalho et al., 2019; Molnar, 2022). Interpretability can also be classified according to its scope. In simple terms, this criterion can be understood as follows. If the interpretability method delivers explanations for understanding the model, it is a global interpretability. On the other hand, if the provided explanation is directly related to a specific prediction, then the method is considered to have local interpretability. A method can also be considered locally scoped if it explains specific predictions for a set of instances (D. V. Carvalho et al., 2019; Molnar, 2022).

AI has been applied in various relevant domains, including healthcare (Schuch et al., 2023; Zhang et al., 2023), finance (Ozbayoglu et al., 2020), criminal justice (Mandalapu et al., 2023), and education (X. Liu et al., 2025; Queiroga et al., 2024). Within the educational domain, this growing use of AI is particularly evident in the field of Educational Data Mining (EDM), which explores statistical, Machine Learning (ML), and Data Mining (DM) algorithms applied to different types of educational data (Romero & Ventura, 2010), through a cyclical knowledge discovery process (Romero & Ventura, 2020), standing out as one of the main forms of big data analysis in education (Xiao et al., 2022). EDM is the research area dedicated to developing methods for discovering insights from various types of data derived from educational environments and utilizing these methods to better understand students and the contexts in which they learn (R. S. J. d. Baker, 2010).

The methods used in EDM are similar to those in general data mining, with multiple approaches available for various applications. Among the most commonly used methods are clustering, association rule mining, regression, and classification (Bakhshinategh et al., 2018; Romero & Ventura, 2010). Applications in the EDM field are diverse and can be categorized based on various criteria (R. S. Baker & Yacef, 2009; Hegazi & Abugroon, 2016; Peña-Ayala, 2014; Romero & Ventura, 2010), including the end user, such as students, educators, administrators, and researchers (Romero & Ventura, 2013). Performance prediction is among the most relevant applications, enabling early identification of the risk of failure or dropout in the learning process (Bakhshinategh et al., 2018; Xiao et al., 2022), in a context where dropout in higher education is an international issue that represents social, academic, and economic waste (R. L. L. Silva Filho et al., 2007).

In this context, this work aims to characterize the state of the art regarding the interpretability of models in Educational Data Mining (EDM). To this end, it proposes a Systematic Literature Mapping (SLM) to identify the key aspects related to (i) interpretability context, (ii) interpretability methods, metrics, and objectives, (iii) education levels, (iv) educational data, and (v) ML techniques. More specifically, beyond mapping these aspects, this study contributes by providing insights into the practical applicability and theoretical implications of interpretability methods in EDM, as follows:

- By relating interpretability methods to practical EDM applications, data types, and education levels, this study provides a comprehensive view of how interpretability is effectively addressed in different educational contexts, highlighting cases where interpretability is treated as a core objective and where it remains secondary;
- By synthesizing studies that measure, compare, or analyze the interpretability of different models and techniques, including both intrinsically interpretable and post hoc approaches, this work identifies best practices and limitations, pointing to the need for systematic comparison and the development of interpretability metrics, particularly for agnostic methods;

- By examining the objectives and motivations behind the use of interpretability methods, this study reinforces the importance of stakeholder involvement and discusses how interpretability interacts with other desiderata in EDM, such as fairness, thereby contributing to the design of transparent, accountable, and context-aware educational systems.

The paper is organized as follows. Section 2 describes another review related to Interpretability. Continuing, Section 3 presents the Research Protocol (RP) for this systematic mapping, including the planning and conduction phases. Section 4 presents the results to answer the proposed research questions. Lastly, Section 5 presents the discussion and future directions, while Section 6 presents the final considerations.

2 Related Work

This section presents some secondary studies that have explored the theme of interpretability, comprising eight non-systematic reviews (nSLR) and seven systematic literature reviews (SLR). None of them, however, focused on interpretability analysis in the context of Educational Data Mining, as this mapping proposes and as highlighted at the end of the section.

The review of Linardatos et al. (2021) serves as a reference point for both theorists and practitioners, as it presents a taxonomy of ML interpretability methods and the best use cases for each approach. In this context, healthcare, finance, computer vision, and natural language processing are among the most frequently discussed domains. Under the proposed taxonomy, four major categories were identified: (i) methods for explaining complex black-box models, (ii) methods for creating white-box models, (iii) methods that promote fairness and mitigate discrimination, and (iv) methods for analyzing the sensitivity of model predictions. The study highlights the significant growth of the sensitivity analysis field, the lack of combinatorial approaches for explaining deep neural networks, and the absence of formal metrics to assess the performance of interpretability methods. Finally, the findings show that despite numerous academic studies on interpretability, these techniques are rarely a significant part of machine learning workflows.

Another study is presented by D. V. Carvalho et al. (2019), covering machine learning interpretability with a focus on methods and metrics. That is an extensive survey with an excellent theoretical foundation for interpretability. Properties and human-friendly characteristics for individual explanations are presented. Three main levels of experiments are proposed for evaluating interpretability. Quantitative and qualitative indicators for interpretability are also presented. Finally, it's highlighted that this research field needs to focus more on comparing existing explanation methods instead of just creating new ones.

The study by Langer et al. (2021) reinforces and extends the focus on human stakeholders as well as on the development and evaluation of explainability approaches, proposing that when tackling explainability, research needs to pay more attention to stakeholders' specific desiderata, such as interests, goals, expectations, needs, and demands regarding AI-based systems. It introduces a conceptual model that explicitly relates the satisfaction of stakeholders' desiderata with the explainability concepts (i) explainability approach, (ii) explanatory information, (iii) stakeholders' understanding, (iv) desiderata satisfaction, and (v) context. Five classes of stakeholders are considered: users, developers, affected parties, deployers, and regulators. Additionally, the review

identifies various desiderata, including acceptance, effectiveness, fairness, and legal compliance, among others. It concludes that, as AI-based systems increasingly influence decision-making processes and societal contexts, stakeholders' desiderata continue to expand. While XAI research has shifted toward human stakeholders, it still lacks a comprehensive view of all stakeholder needs in socially relevant contexts.

The study by Khosravi et al. (2022) makes contributions by addressing the specific challenges and mechanisms of XAI in education, proposing an XAI in Education (XAI-ED) framework that draws on the fields of AI, Human-Computer Interaction, and the Cognitive and Learning Sciences. Regarding stakeholders, for example, it highlights that providing explanations solely for system engineers or data scientists is not sufficient, as human factors are essential for the effective communication of explanations in any context. Understanding and trust in AI depend on the people involved. Additionally, four case studies are presented to illustrate the application of XAI-ED, including an adaptive learning system, feedback within a writing analytics tool, and data narratives through teamwork analysis in educational healthcare systems.

The Rachha and Seyam (2023) work addressed the Explainable Artificial Intelligence (XAI) in education. One of its contributions is summarizing the explainability techniques used in various domains, identifying gaps, and considering their potential usefulness in education. The survey presents some discussions as follows: the need for a unified framework for XAI in Education; the need for established principles and guidelines for Human-in-the-loop XAI systems; the need for a holistic consideration of incorporating multiple disciplines in XAI design; the need for Identification of associated risk and mitigation; future directions for educators, policymakers, researchers, and developers.

Yet in the nSLR group, there are brief reviews. According to S. Chen et al. (2023), the practical implementation of EDM in the context of EDM has fallen short of desired outcomes due to challenges such as data fragmentation, a lack of interpretability in analyses, and usability issues. In this regard, it proposes a collaboration-centric strategy for EDM application design, encompassing improvements at three levels: a unified data platform that aggregates multi-source data, interpretable data analysis middleware, and customizable front-end services tailored to multiple roles. Pantazatos et al. (2024) presents a brief review that proposes the application of XAI in the context of network management education, enhancing the learning experience by improving transparency and empowering instructors and students in the field of IT infrastructure management. A discussion about the means and applications of XAI in various fields is presented, including education. Tousside et al. (2022) reviews the current state of explainability in modern EDM approaches, including discussion about the advantages of its use, as well as the EDM Approaches addressing or lacking explainability. Additionally, propose some XAI methods that could be incorporated into specific state-of-the-art EDM approaches.

The following reviews are systematic literature reviews, and since their methodology is closer to the mapping study proposed here, we present Table 1 to highlight (i) their base search string and (ii) whether the review focuses on Education. Next, each of these reviews is described individually.

The Vieira and Digiampietri (2022) review focused on post-hoc interpretability in Machine Learning, independent of the research area. In this regard, certain restrictions were applied during the selection process. The papers must use tabular and labeled data and propose a model-agnostic

Table 1: Related work identified as Systematic Literature Reviews.

Work	Search String Scope	Focus on Education
Vieira and Digiampietri (2022)	Title, abstract, and keyword fields. ((“black-box” OR “black box”) AND (“machine learning” OR “predict*” OR “model*” OR “classif*”) AND (“explainable” OR “explainability” OR “interpretability”) AND (“agnostic” OR “post-hoc” OR “post hoc”))	No
Araujo (2021)	Title and abstract fields. (framework of interpretability of machine learning”) OR (“tool of interpretability in machine learning”) OR (“tools of explainability of machine learning models”) OR (“framework of explainability of machine learning”) OR (“interpretable machine learning tool”) OR (“interpretable machine learning”) OR (“explainable machine learning tools”) OR (“explainable frameworks of machine learning”) OR (“explainability of machine learning models”) OR (“performance of framework of interpretability machine learning”) OR (“machine learning interpretability techniques”) OR (“decipherable machine learning”) OR (“understandable machine learning”) OR (“explicable machine learning”)	No
Deck et al. (2024)	(xai OR explanation OR understandab* OR intelligib* OR comprehensib* OR interpretab* OR explainab* OR transparen*) AND fair* AND (ai OR “artificial intelligence” OR “machine learning”)	No
ŞAHiN et al. (2025)	Web of Science (TI=(explain* deep learning or interpre* deep learning or reliab* deep learning)) OR AK=(explain* deep learning or interpre* deep learning or reliab* deep learning)	No
Alamri and Alharbi (2021)	Example for Web of Science TI=(student* OR academic) AND (predict* OR model*) AND (performance OR grade) TS=(white-box OR Interpretable OR understandable OR explainable OR expressiviness OR Rule*)	Yes*
Raji et al. (2023)	- ML AND student performance OR EDM AND student performance - Academic performance AND Deaf - ML AND students with Hearing Impairment OR EDM AND students with Hearing Impairment AND student performance prediction - Open Dataset AND student performance	Yes**
Pandian et al. (2024)	Keywords such as "Educational Data Mining", "Student Performance Prediction", "Evaluations of Students", "Performance Analysis of Students", and "Learning Curve Prediction"	Yes*

Legend: Yes* - student performance prediction, Yes** - academic performance of deaf children.

approach. The goal is to identify the methods used for interpretability, their metrics, and the challenges still to be addressed in the field. The searches were performed in June 2020, and after the selection phase, 11 studies were accepted for analysis. According to the review, determining the correct measurement criteria and metric for each case is challenging and remains an open problem in the field.

The study by Araujo (2021) reviewed the use of interpretability frameworks in machine learning, aiming to identify the most frequently adopted approaches, the ML algorithms used, and the domains in which they were applied. The authors also sought to highlight research related to the performance of these interpretability techniques. Among the analyzed frameworks, LIME stood out with 19 occurrences, followed by SHAP with 3 occurrences. Regarding the application domains, the most common areas were Health, Natural Language Processing (NLP), and Finance. Notably, none of the selected works addressed interpretability in the context of Education. The

review analyzed 26 studies published between 2017 and 2020, retrieved from IEEE[1], ACM Digital Library[2], and Science Direct[3].

Deck et al. (2024) investigate the different relationships between explainability and algorithmic fairness. The paper organizes and critically analyzes the claims found in recent literature regarding the benefits of Explainable Artificial Intelligence (XAI) for fairness, including its role in analyzing, reporting, and mitigating unfairness. From the perspective of domain application, it provides a general review and aligns with stakeholders' desiderata, as noted by Langer et al. (2021). The study by ŞAHiN et al. (2025) provides a comprehensive overview of state-of-the-art methods and techniques for enhancing interpretability, explainability, and reliability in deep learning models. Among its contributions are the proposal of a taxonomy of XAI methods and the exploration of domain-specific applications of XAI methods, demonstrating how these techniques are tailored to meet the unique requirements of diverse fields.

The study by Alamri and Alharbi (2021) presented a systematic review considering articles that utilize explainable ML models in student performance prediction. The goal was to identify the performance measures to be predicted, the predictors used to train the models, and the methods used. Also, it aimed to identify evaluation metrics to assess the explainability of the models and the methods that meet both requirements of high accuracy and explainability. As a result, none of the selected studies has utilized any evaluation metric to assess the explainability of the models, revealing a critical shortcoming of existing research in explainable student performance prediction models. The search was applied in ISI Web of Science[4] and Google Scholar[5], considering the period from 2015 to 2020.

The review Raji et al. (2023) aims to shed light on the importance of research in the field of deaf education utilizing ML techniques. Exploring the advantages of using explainable models in this domain identifies that they are crucial for decision-making and for building stakeholders' trust in the educational system. The study Pandian et al. (2024) conducts an exhaustive review of previous research on the application of ML algorithms to predict the academic performance of students in various educational environments. Although interpretability is not explicitly addressed in the research questions, the study reveals that only 6% of the concerns identified in the limitations of existing algorithms are related to interpretability.

The research proposed here differs from previous studies in several aspects: (i) it is a Systematic Literature Mapping (SLM) focused on interpretability in Educational Data Mining (EDM), (ii) there are no restrictions on the type or scope of interpretability methods, (iii) there are no restrictions on the EDM application, (iv) there are no restrictions on the type of data used, (v) it includes a research question to assess initiatives aimed at quantifying model interpretability, (vi) it includes a research question to identify the objectives behind the use of interpretability, and (vii) it verifies whether studies attempt to compare the interpretability of different models. Thus, this review presents a relevant contribution to EDM in the field of explainability.

[1] http://ieeexplore.ieee.org
[2] http://portal.acm.org
[3] https://www.sciencedirect.com/
[4] http://www.isiknowledge.com
[5] https://scholar.google.com/

3 Methodology

The present Section describes the Research Protocol (RP) for the proposed Systematic Literature Mapping (SLM). This protocol is based on guidelines from Kitchenham and Charters (2007), which is also widely used in Informatics in Education (Dermeval et al., 2020). The planning and conduction phases were carried out with the support of the online tool Parsifal[6].

3.1 Research Questions

The Research Questions (RQ) are the first step in the SLM. Aiming to characterize the state of the art in the themes of Model Interpretability and EDM, the following Research Questions were defined:

RQ1 – In which context was interpretability explored?

RQ2 – In which level of education were the studies applied?

RQ3 – What educational data were used?

RQ4 – What methods were used to obtain the models in the context of the EDM?

RQ5 – Which strategy (method or framework) was used to analyze the interpretability of the models?

RQ6 – What metrics were used to assess the interpretability of a model?

RQ7 – What is the purpose or objective when addressing interpretability?

3.2 Digital Libraries and Search String

Primary studies were searched in ACM Digital Library[7], IEEE Digital Library[8], ISI Web of Science[9], Scopus[10], and SBC-OpenLib (SOL)[11]. These scientific bases enable advanced search through "search strings". These search strings are expressions that combine words of interest and logical operators.

In that regard, the elaborated string uses terms related to EDM, Interpretability, and Explainability. Table 2 presents the final search string for each Digital Library. The desired terms are defined from these expressions, where they should be found in the paper (Title, Abstract, or Keywords), and finally, the Year of publication. This mapping study, therefore, is restricted to studies published from 2017 onwards[12]. The search string for Scopus can be considered the base string, as its format is concise and easy to understand.

[6]https://parsif.al/
[7]http://portal.acm.org
[8]http://ieeexplore.ieee.org
[9]http://www.isiknowledge.com
[10]http://www.scopus.com
[11]https://sol.sbc.org.br/
[12]In the IEEE search, the Year of publication is manually filtered after the search string is applied.

Table 2: Search Strings for each Digital Library.

Digital Library	Search String
ACM Digital Library	(Title:(“educational data mining") OR Abstract:(“educational data mining") OR Keyword:(“educational data mining")) AND (Title:(“explainable") OR Title:(“explainability") OR Title:(“interpretability") OR Title:(“interpretable") OR Title:(“explanation*") OR Abstract:(“explainable") OR Abstract:(“explainability") OR Abstract:(“interpretability") OR Abstract:(“interpretable") OR Abstract:(“explanation*") OR Keyword:(“explainable") OR Keyword:(“explainability") OR Keyword:(“interpretability") OR Keyword:(“interpretable") OR Keyword:(“explanation*")) AND [Publication Date: (01/01/2017 TO 12/31/2024)]
IEEE Digital Library	((“Document Title":“educational data mining") OR (“Abstract":“educational data mining") OR (“Author Keywords":“educational data mining")) AND (“Document Title":explainable OR “Document Title":explainability OR “Document Title":interpretability OR “Document Title":interpretable OR “Document Title":explanation* OR “Abstract":explainable OR “Abstract":explainability OR “Abstract":interpretability OR “Abstract":interpretable OR “Abstract":explanation* OR “Author Keywords":explainable OR “Author Keywords":explainability OR “Author Keywords":interpretability OR “Author Keywords":interpretable OR “Author Keywords":explanation*)
ISI Web of Science	(TI=("educational data mining") OR AB=("educational data mining") OR AK=("educational data mining")) AND (TI=(explainable OR explainability OR interpretability OR interpretable OR explanation*) OR AB=(explainable OR explainability OR interpretability OR interpretable OR explanation*) OR AK=(explainable OR explainability OR interpretability OR interpretable OR explanation*)) AND DOP=(2017-01-01/2024-12-31)
Scopus	TITLE-ABS-KEY ((explainable OR explainability OR interpretability OR interpretable OR explanation*) AND (“educational data mining")) AND PUBYEAR $>$ 2016 AND PUBYEAR $<$ 2025
SBC-OpenLib (English and Portuguese)	All ((explainable OR explainability OR interpretability OR interpretable OR explanation*) AND (educational data mining)) AND PUBYEAR $>$ 2016 AND PUBYEAR $<$ 2025, All ((explicável OR explicabilidade OR interpretabilidade OR interpretável OR explicação OR explicações) E (mineração de dados educacionais)) AND PUBYEAR $>$ 2016 AND PUBYEAR $<$ 2025

3.3 Inclusion and Exclusion criteria

The next activity in the planning phase is to define Inclusion (IC) and Exclusion Criteria (EC), which will be applied to all studies obtained from applying the search string. The Exclusion Criteria are as follows: Duplicate papers (EC1); Short papers (EC2); Papers written in a language other than English or Portuguese[13] (EC3); Secondary studies (EC4); Books, technical reports, and other forms of gray literature (EC5); Papers that do not address interpretability in the study or the context of EDM (EC6); Paper not available (EC7). Moreover, the Inclusion Criterion is: Addresses aspects of interpretability in the context of EDM (IC1).

[13]Only when SBC OpenLib is being consulted.

3.4 Conduction

Once planned, the mapping moves on to the conduction phase. The plan defined in the previous Sections is applied, and a set of studies is obtained. Finally, data are extracted from these studies to respond to the proposed Research Questions. The conduction process was structured into several stages. The first stage involves applying search strings to all scientific databases. This query is up-to-date until December 2024. At this time, applying some exclusion criteria independently of paper reading is possible. More specifically the EC1, EC2, EC3, EC5 and EC7 criteria.

In the second stage, the inclusion and exclusion criteria are applied, considering the Title and Abstract of the papers. Lastly, in the third stage, the same criteria are used, considering the full text of the documents. Table 3 presents the number of papers discarded for each exclusion criterion and scientific base, and shows that 65 papers were obtained at the end of the third stage. Table 4 presents this final list of documents. Half of the publications came from six countries – China (12), the USA (6), Brazil (5), India (5), Canada (4), and Spain (4) – with China alone representing 18% of all studies.

Table 3: Papers discarded during the conduction phase of SLM.

Base	Initial search	EC1	EC2	EC3	EC4	EC5	EC6	EC7	Final
ACM	15	13	0	0	0	0	2	0	0
IEEE	42	38	2	0	0	0	1	0	1
ISI WoS	89	86	0	0	0	0	1	1	1
Scopus	165	5	7	1	11	8	60	13	60
SOL	4	1	0	0	0	0	0	0	3
	315	143	9	1	11	8	64	14	65

Table 4: Final set of selected papers.

#	Citation	Title
1	Tsiakmaki and Ragos (2021)	A Case Study of Interpretable Counterfactual Explanations for the Task of Predicting Student Academic Performance
2	Chau and Phung (2021)	A Cumulative Increasing Kemelized Nearest-Neighbor Bagging Method for Early Course-Level Study Performance Prediction
3	R. L. C. Silva Filho et al. (2023)	A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement
4	J. Z. Wang et al. (2017)	A latent factor model for instructor content preference analysis
5	Suaza-Medina et al. (2024)	A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations
6	Hooshyar et al. (2022)	A Three-Layered Student Learning Model for Prediction of Failure Risk in Online Learning
7	Lu et al. (2024)	Advanced Knowledge Tracing: Incorporating Process Data and Curricula Information via an Attention-Based Framework for Accuracy and Interpretability
8	Livieris et al. (2023)	An Advanced Explainable and Interpretable ML-Based Framework for Educational Data Mining
9	Gámez-Granados et al. (2023)	An algorithm based on fuzzy ordinal classification to predict students' academic performance
10	Rangone et al. (2022)	An Educational Data Mining Model based on Auto Machine Learning and Interpretable Machine Learning

#	Citation	Title
11	Cavus and Kuzilek (2024)	An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations of Student Success Prediction Models
12	Alwarthan et al. (2022)	An Explainable Model for Identifying At-Risk Student at Higher Education
13	Pei and Xing (2022)	An Interpretable Pipeline for Identifying At-Risk Students
14	Choi et al. (2024)	Analyzing the Interpretability of Machine Learning Prediction on Student Performance Using SHapley Additive exPlanations
15	Chou (2023)	Apply an Integrated Responsible AI Framework to Sustain the Assessment of Learning Effectiveness
16	Chou (2021)	Apply explainable AI to sustain the assessment of learning effectiveness
17	Novillo Rangone et al. (2022)	Automation of an Educational Data Mining Model Applying Interpretable Machine Learning and Auto Machine Learning
18	Alharbi (2022)	Back to Basics: An Interpretable Multi-Class Grade Prediction Framework
19	Zanellati et al. (2024)	Balancing Performance and Explainability in Academic Dropout Prediction
20	Qu et al. (2022)	Can We Predict Student Performance Based on Tabular and Textual Data?
21	Joshi et al. (2021)	CatBoost - An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance
22	Lallé et al. (2021)	Combining data-driven models and expert knowledge for personalized support to foster computational thinking skills
23	Arévalo-Cordovilla and Peña (2024)	Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors
24	Aytekin and Saygín (2025)	Discovering prerequisite relations using large language models
25	Pu et al. (2022)	Embedding cognitive framework with self-attention for interpretable knowledge tracing
26	Parkavi et al. (2024)	Enhancing personalized learning with explainable AI: A chaotic particle swarm optimization-based decision support system
27	Islam et al. (2024)	Enhancing tertiary students' programming skills with an explainable Educational Data Mining approach
28	Thuy and Benoit (2024)	Explainability through uncertainty: Trustworthy decision-making with neural networks
29	Guleria and Sood (2023)	Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling
30	Q. Liu and Khalil (2024)	Explainable AI in Learning Analytics: Improving Predictive Models and Advancing Transparency Trust
31	Alonso and Casalino (2019)	Explainable Artificial Intelligence for Human-Centric Data Analysis in Virtual Learning Environments
32	Ghimire et al. (2024)	Explainable artificial intelligence-machine learning models to estimate overall scores in a tertiary preparatory general science course
33	Abdalkareem and Min-Allah (2024)	Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia
34	F. Liu et al. (2024)	FDKT: Towards an Interpretable Deep Knowledge Tracing via Fuzzy Reasoning
35	Luo and Wang (2024)	Feature Mining Algorithm for Student Academic Prediction Based on Interpretable Deep Neural Network
36	Q. Liu et al. (2018)	Fuzzy cognitive diagnosis for modelling examinee performance
37	Lemay and Doleck (2020)	Grade prediction of weekly assignments in MOOCS: mining video-viewing behavior
38	Nakagawa et al. (2021)	Graph-based knowledge tracing: Modeling student proficiency using graph neural network
39	Lee (2023)	Identifying Prerequisite Courses in Undergraduate Biology Using Machine Learning
40	Vultureanu-Albişi and Bădică (2021)	Improving Students' Performance by Interpretable Explanations using Ensemble Tree-Based Approaches
41	Huang et al. (2024)	Interpretable neuro-cognitive diagnostic approach incorporating multidimensional features

#	Citation	Title
42	Lu et al. (2022)	Interpreting Deep Learning Models for Knowledge Tracing
43	F. Liu et al. (2023)	Interpreting Learner Success: Enhancing Knowledge Tracing with Attention-Based IRT Models in Modern Education
44	Colpo et al. (2024)	Lessons learned from the student dropout patterns on COVID-19 pandemic: An analysis supported by machine learning
45	F. Chen and Cui (2020)	LogCF: Deep collaborative filtering with process data for enhanced learning outcome modeling
46	Matetic (2019)	Mining learning management system data using interpretable neural networks
47	Gupta et al. (2022)	Mining Sequential Learning Trajectories with Hidden Markov Models for Early Prediction of At-Risk Students in E-Learning Environments
48	Shen et al. (2023)	Monitoring Student Progress for Learning Process-Consistent Knowledge Tracing
49	Y. Wang et al. (2023)	National student loans default risk prediction: A heterogeneous ensemble learning approach and the SHAP method
50	Geng et al. (2023)	Noise-Filtering Enhanced Deep Cognitive Diagnosis Model for Latent Skill Discovering
51	Van Petegem et al. (2023)	Pass/Fail Prediction in Programming Courses
52	Jang et al. (2022)	Practical early prediction of students' performance using machine learning and eXplainable AI
53	Kumar and Sharma (2020)	Predicting Academic Performance of International Students Using Machine Learning Techniques and Human Interpretable Explanations Using LIME—Case Study of an Indian University
54	Singelmann et al. (2020)	Predicting and Understanding Success in an Innovation-Based Learning Course
55	Lemay and Doleck (2022)	Predicting completion of massive open online course (MOOC) assignments from video viewing behavior
56	Nnadi et al. (2024)	Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models
57	Al-Jallad et al. (2019)	Rule mining models for predicting dropout/ stopout and switcher at college using satisfaction and SES features
58	Colak Oz et al. (2023)	School dropout prediction and feature importance exploration in Malawi using household panel data: machine learning approach
59	Jeon et al. (2019)	Time-series insights into the process of passing or failing online university courses using neural-induced interpretable student states
60	Cohausz (2022)	Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science
61	Alvarez-Garcia et al. (2024)	Uncovering student profiles. An explainable cluster analysis approach to PISA 2022
62	Hooper et al. (2023)	Using Machine Learning in Veterinary Medical Education: An Introduction for Veterinary Medicine Educators
63	C. Carvalho et al. (2023)	Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior
64	Gusmão et al. (2021)	A qualidade da educação para além do IDEB: Um estudo através de técnicas de Mineração de Dados
65	Gama Neto et al. (2021)	Mineração de Dados Aplicada à Predição do Desempenho de Escolas e Técnicas de Interpretabilidade dos Modelos

4 Report

This Section presents the SLM results, organized to answer the proposed research questions. Discussion regarding the findings and possible future work is addressed in Section 5. At the end of some of the following sections, we present a few examples of practical implications. Although not intended to be exhaustive, these cases highlight both strengths and limitations when addressing interpretability in EDM.

4.1 Interpretability context

This topic is related to the RQ1 question "In which context was interpretability explored?" and the objective is to group selected works according to the problem they propose to solve.

Among the selected studies, performance prediction was the most reported application, with forty-seven (72.3%) occurrences. In these cases, while some studies focus on identifying students at risk of dropping out or failing, others classify students into different performance levels or predict their grades in courses or exams. There are also cases of predicting school performance in qualifying exams. Other categories were Knowledge Tracing (7), Cognitive Modeling (2), Academic Pathways (1), Adaptability prediction (1), Career Counseling (1), Discovery of student profiles (1), Instructor Preference (1), Latent Skill (1), Open Problem (1), Prerequisites (1), and Student loan default (1). Table 5 identifies the interpretability context of each study. All cases are discussed in the following.

Table 5: List of papers according to the interpretability context. The ID is based on Table 4.

Interpretability context	Papers IDs	Total
Performance Prediction	1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 26, 27, 28, 30, 31, 32, 35, 37, 39, 40, 44, 45, 46, 47, 51, 52, 53, 54, 55, 57, 58, 59, 60, 62, 63, 64, and 65	47
Knowledge Tracing	7, 25, 34, 38, 42, 43, and 48	7
Cognitive Modeling	36, and 41	2
Academic Pathways	33	1
Adaptability prediction	56	1
Career counseling	29	1
Discovery of student profiles	61	1
Instructor Preference	4	1
Latent Skill	50	1
Open problem	17	1
Prerequisites	24	1
Student loan default	49	1

Some works presented particular contexts. The work by Guleria and Sood (2023) explored the area of career counseling, while the study by J. Z. Wang et al. (2017) addressed instructors' preferences when excluding questions from student assignments. Cognitive modeling of examinees (Huang et al., 2024; Q. Liu et al., 2018) and Latent skill (Geng et al., 2023) were also addressed. Another aspect was Knowledge tracing, considered by the studies F. Liu et al. (2024),

F. Liu et al. (2023), Lu et al. (2024), Lu et al. (2022), Nakagawa et al. (2021), Pu et al. (2022), and Shen et al. (2023).

Personalized support to foster computational thinking skills was addressed by Lallé et al. (2021), and prediction in programming learning was part of the studies by Arévalo-Cordovilla and Peña (2024), Choi et al. (2024), and Islam et al. (2024). Meanwhile, student success in innovation-based learning, with training for engineers, was analyzed by Singelmann et al. (2020). Problems such as predicting academic paths, including those in computer science and engineering, health, business, religion, and general science, were addressed by Abdalkareem and Min-Allah (2024). Meanwhile, student classification into adaptability levels is presented in Nnadi et al. (2024).

The use of academic data was also useful for predicting the risk of student loan default, as reported by Y. Wang et al. (2023). In the context of course curricula, Aytekin and Saygín (2025) explored the automatic detection of prerequisite relationships between concepts. The discovery of student profiles was addressed by Alvarez-Garcia et al. (2024), while models for various problems, including dropout, procrastination, breakdown, and student performance, were presented by Rangone et al. (2022).

Examples of student performance prediction are the studies by Cavus and Kuzilek (2024), Chau and Phung (2021), and Tsiakmaki and Ragos (2021). While R. L. C. Silva Filho et al. (2023) proposes a framework to reveal relevant contextual features for predicting educational outcomes, Qu et al. (2022) utilizes tabular and textual data. Prediction in MOOC courses was addressed by Lemay and Doleck (2020, 2022) and Thuy and Benoit (2024), while performance classification into multiple categories can be found in the works of Joshi et al. (2021) and Vultureanu-Albişi and Bădică (2021). Grade prediction for the next academic period was also a case study in Alharbi (2022).

Student performance prediction was also studied in a general science preparatory course for higher education (Ghimire et al., 2024), in the prediction of success in obtaining a degree in biology (Lee, 2023), and in predicting student performance in a veterinary course. Student performance in underdeveloped regions was addressed by Suaza-Medina et al. (2024). Other studies focusing on student performance prediction are presented by Alonso and Casalino (2019), F. Chen and Cui (2020), Chou (2021), Gámez-Granados et al. (2023), Jeon et al. (2019), Kumar and Sharma (2020), Q. Liu and Khalil (2024), Livieris et al. (2023), Luo and Wang (2024), Matetic (2019), and Parkavi et al. (2024).

Colpo et al. (2024) addressed performance prediction by identifying the main dropout patterns and how these patterns compared before and during the COVID-19 pandemic. Studies focused on at-risk or dropout prediction were also conducted by Al-Jallad et al. (2019), Alwarthan et al. (2022), Chou (2023), Cohausz (2022), Colak Oz et al. (2023), Colpo et al. (2024), Gupta et al. (2022), Hooshyar et al. (2022), Jang et al. (2022), Pei and Xing (2022), Van Petegem et al. (2023), and Zanellati et al. (2024). Finally, we can highlight two cases of school performance prediction in Brazil, as in Gama Neto et al. (2021) regarding the São Paulo State School Performance Assessment System (SARESP) and in Gusmão et al. (2021) regarding the Basic Education Development Index (IDEB).

4.2 Level of Education

This topic is related to the RQ2 question "In which level of education were the studies applied?". To answer this question, we tried to classify the studies into basic education, higher education, and postgraduate categories. In situations where the level of education is not explicitly stated in the article text, an attempt was made to make this deduction based on information from the dataset itself.

The analysis shows that Undergraduate courses represent the most explored level of education in the selected studies. Forty (40) studies examined Undergraduate courses, either alone or in combination with another level such as Graduate courses, corresponding to 61.5% of the studies. Basic Education appears in twenty-two (22) studies, indicating a consistent presence across EDM applications. In contrast, Graduate courses appear in only a few cases, either in isolation or combined with other levels, demonstrating limited exploration at this level.

To better understand how educational levels are distributed across Interpretability contexts, Table 6 relates Level of education to Interpretability context. The results indicate that Performance Prediction — the predominant context — is mainly associated with Undergraduate courses, although it also includes applications in Basic Education and in combined levels. Knowledge Tracing frequently involves Basic Education, either exclusively or in combination with Undergraduate courses, while other contexts such as Cognitive Modeling, Academic Pathways, Adaptability prediction, Discovery of student profiles, and Latent Skill are also restricted to Basic Education. Contexts such as Prerequisites and Student loan default appear exclusively in Undergraduate courses, and Career counseling is classified under All levels.

Applications in EDM are strongly concentrated in Undergraduate courses within the Performance Prediction context, whereas Basic Education is more represented in Knowledge Tracing and other specialized contexts. Graduate courses remain marginal in the analyzed literature.

Table 6: Relationship between interpretability context and the level of education. The ID is based on Table 4.

Interpretability context	**Level of education**	**Papers IDs**	Total
Performance Prediction	Undergraduate courses	1, 2, 6, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 23, 26, 27, 28, 31, 37, 44, 39, 46, 47, 51, 54, 55, 57, 60, 62, and 63	47
	Basic Education	3, 5, 21, 22, 35, 40, 45, 58, 64, and 65	
	Others	32	
	Undergraduate and Graduate courses	52, 53, and 59	
	Basic Education and Undergraduate courses	8	
	Graduate courses	30	
Knowledge Tracing	Basic Education	7, 34, and 48	7
	Others	25, 38, and 42	
	Basic Education and Undergraduate courses	43	
Cognitive Modeling	Basic Education	36 and 41	2
Academic Pathways	Basic Education	33	1
Adaptability prediction	Basic Education	56	1
Career counseling	All levels	29	1
Discovery of student profiles	Basic Education	61	1
Instructor Preference	Others	4	1
Latent Skill	Basic Education	50	1
Open problem	Undergraduate and Graduate courses	17	1
Prerequisites	Undergraduate courses	24	1
Student loan default	Undergraduate courses	49	1

4.3 Educational Data

This topic is related to the RQ3 question "What educational data were used?". The purpose is to identify and list the primary sources used in the studies and group papers based on these findings. Virtual learning environments, academic record systems, and surveys are among the identified sources. Table 7 presents the relationship between the educational data, the interpretability context, and the level of education.

Learning Management Systems (LMS) and other specialized platforms used for online courses or training stand out in the virtual environment category. In these environments, the data are diverse and may include behavior patterns such as interactions in discussion forums, quiz attempts, task submissions, reading, and consumption of resources, including videos.

Academic management systems or academic records also allow for the retrieval of various information. The first group includes the student's academic performance data, which is related, for example, to grades, averages, and previous achievements. These systems, in a complementary way, often also store demographic and socio-economic information about the student.

Another identified source is research or surveys, which enable the collection of well-targeted data for research. This group may also include academic performance, demographic, and socio-economic information. Next, the studies are analyzed with regard to this topic.

Table 7: Relationship between educational data, interpretability context, and level of education. The ID is based on Table 4.

Educational data	**Interpretability context**	**Level of education**	**Papers IDs**	Total
Virtual environments	Performance Prediction	Undergraduate courses	1, 6, 9, 11, 15, 16, 28, 31, 37, 46, 47, 51, 54, and 55	14
	Performance Prediction	Graduate courses	30	1
	Performance Prediction	Undergraduate and Graduate courses	52 and 59	2
	Performance Prediction	Basic Education	22 and 45	2
	Knowledge Tracing	Basic Education	7, 34, and 48	3
	Knowledge Tracing	Basic Education and Undergraduate courses	43	1
	Instructor Preference	Others	4	1
Academic records	Performance Prediction	Undergraduate courses	10, 18, 19, 27, 39, 57, 60, 62, and 63	9
	Performance Prediction	Basic Education	3	1
	Performance Prediction	Undergraduate and Graduate courses	53	1
	Student loan default	Undergraduate courses	49	1
	Academic Pathways	Basic Education	33	1
	Adaptability prediction	Basic Education	56	1
Virtual environments and academic records	Performance Prediction	Undergraduate courses	2, 12, 13, 14, 20, 23, and 44	7
	Performance Prediction	Basic Education	21 and 40	2
	Performance Prediction	Basic Education and Undergraduate courses	8	1
Answering exams	Performance Prediction	Basic Education	5, 64, and 65	3
	Performance Prediction	Others	32	1
	Cognitive Modeling	Basic Education	36	1
	Latent Skill	Basic Education	50	1
	Discovery of student profiles	Basic Education	61	1
	Knowledge Tracing	Others	25, 38, and 42	3
Questionnaires	Performance Prediction	Basic Education	35	1
	Performance Prediction	Undergraduate courses	26	1
	Cognitive Modeling	Basic Education	41	1
Dynamic (including survey)	Open problem	Undergraduate and Graduate courses	17	1
Household data	Performance Prediction	Basic Education	58	1
Placement data	Career counseling	All levels	29	1
Prerequisites	Prerequisites	Undergraduate courses	24	1

In the work group that only considered data from virtual environments, the study by Gupta et al. (2022) used the OULA public dataset, which has a sufficient number of LWAs and 22 modules from STEM and Social Sciences. The target attribute was mapped to at-risk and safe classes, obtaining a binary classification problem.

Hooshyar et al. (2022) used LMS data, which was classified into content access, engagement, and review behavior. Examples of content access include course views, resource views, downloads, and URL views, among others. Regarding engagement, we can mention viewing dis-

cussions in forums and adding to or commenting on discussions. Finally, regarding behavior in assessments, examples are views or quiz participation.

Jang et al. (2022) carried out online courses using the LMS system (Blackboard), where students could watch videos, download reading materials, submit assignments, participate in active discussions, and take exams. The number of students per course ranged from 65 to 380, totaling 940. Thirteen practical characteristics were considered: 12 related to online behavior and 1 to evaluation.

The two works by Lemay and Doleck (2020) and Lemay and Doleck (2022) utilized data from virtual environments, focusing on student behavior while watching videos. In this case, characteristics such as reproductions, pauses, searches, and changes in the reproduction rate were taken into account.

Tsiakmaki and Ragos (2021) used log data from a one-semester course supported by online resources and activities over the Moodle learning platform. For the case study, a compulsory lecture course, "Physical Chemistry I," was selected and conducted during the spring semester of the 2017-2018 academic year at Aristotle University of Thessaloniki in Greece. A total of 282 students attended this course.

The studies Chou (2021, 2023) conducted experiments using a dataset collected from an LMS system comprising 1,040 anonymized students across various study levels. The data pertains to their online activities and academic results, comprising 16 variables.

Matetic (2019) work with a dataset related to a programming course obtained from Moodle. It contains 408 instances for five generations of students and five attributes: the total number of scores the students received within the lectures, the total number of scores the students received on two quizzes, the total number of scores the students received in labs, the total number of views of the video lectures, target attribute (pass/fail).

The study by J. Z. Wang et al. (2017) collected instructors' preferences for excluding questions from being given to learners in their class via OpenStax Tutor [13], a personalized learning and teaching platform. The OULAD (Open University Learning Analytics Dataset) was used by Cavus and Kuzilek (2024) and Gámez-Granados et al. (2023). This dataset contains information about seven independent courses from different domains. The data used includes demographic information, registration information, assessment information, and Virtual Learning Environment (VLE) log information.

The ASSISTments online learning platform is presented in several studies that focus exclusively on virtual environments, as in F. Liu et al. (2024), F. Liu et al. (2023), and Lu et al. (2024). ASSISTments is widely used at various levels of K-12 education, including information about students' interactions with exercises, knowledge components (skills), and scores.

Student-course records from HarvardX and MITx massive open online courses (MOOCs), encompassing a diverse range of topics, processed clickstream activities, and student demographics, are utilized by Thuy and Benoit (2024). Another MOOC dataset used is from a Norwegian Open edX platform with the help of the learning analytics tool, OXALIC. Q. Liu and Khalil (2024) utilized four categories of data, including course statistics, course interaction-related data, course video-related data, and time spent on the platform.

A custom online learning management system (LMS) was developed for students to track their learning objectives and deliverables. In this case, Singelmann et al. (2020) utilized both quantitative and textual data to represent eight objectives and 32 deliverables for 28 students over a semester.

Other works, not less important, also focused on data from virtual environments (Alonso & Casalino, 2019; F. Chen & Cui, 2020; Jeon et al., 2019; Lallé et al., 2021; Shen et al., 2023; Van Petegem et al., 2023).

Some works combined data from virtual environments with data from academic records, which in practice involve, for example, demographic and socioeconomic data. Some examples are Alwarthan et al. (2022), Chau and Phung (2021), Joshi et al. (2021), Pei and Xing (2022), Qu et al. (2022), and Vultureanu-Albişi and Bădică (2021). Here, the OULAD dataset is also present, as addressed by Choi et al. (2024), and includes demographic data, behavioral metrics, and academic achievements.

Livieris et al. (2023) used two real-world educational datasets. The first includes information about 337 university students who attended an academic course using a Learning Content Management System (LMS) in a blended learning environment. The second contains data about 3,716 students who attended mathematics courses in a secondary school. This dataset summarizes information about the students' performance from the first two out of three semesters, such as test grades, final examination grades, oral grades, and semester grades.

Data from the Moodle Learning Management System (LMS) and external factors of 591 students enrolled in online object-oriented programming courses at the Universidad Estatal de Milagro (UNEMI) were used by Arévalo-Cordovilla and Peña (2024). The target class for at-risk students was defined by a cutoff point in grade scores.

Colpo et al. (2024) used data from 3,371 undergraduate students from the Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar) with distinction for the pre-pandemic period (2019) and the period during the pandemic (2020). The data includes a Virtual Learning Environment (VLE), Academic, contextual, economic, interactional, social, and demographic information.

On the other hand, some studies only considered data from academic records (Alharbi, 2022; Al-Jallad et al., 2019; Kumar & Sharma, 2020; Rangone et al., 2022; R. L. C. Silva Filho et al., 2023). The studies by Geng et al. (2023), Q. Liu et al. (2018), Lu et al. (2022), Nakagawa et al. (2021), and Pu et al. (2022) used exam response data. Student placement data was used by Guleria and Sood (2023). Whereas Colak Oz et al. (2023) used a set of variables obtained from household data.

Some examples using university data can be provided. Zanellati et al. (2024) used a comprehensive dataset that includes demographic information, previous educational metrics, and real-time academic performance indicators. The data refer to the first year of more than 40,000 students from 110 courses of an Italian university. C. Carvalho et al. (2023) utilized a rich dataset comprising over 14,000 students from all undergraduate courses at the Federal University of Pelotas to address the issue of school dropout in higher education. The data represent the pre-pandemic period and include 41 attributes covering demographic information and academic performance.

The dataset used by Islam et al. (2024) comprises 1,720 samples and encompasses data from Computer Science and Engineering (CSE) students attending various universities in Bangladesh. It encompasses a broad spectrum of information, including students' proficiency in programming languages, problem-solving experiences in online and onsite programming contests, computational skills, creativity, course results, and other technological experiences, resulting in 36 features to predict students' proficiency level.

The University of Beijing is another example. In Y. Wang et al. (2023), data attributes include basic student information, personal honors, grades (GPA) by semester, and loan information. The data pertain to students who have applied for national student loans. Artificial and real-life data were also used in the context of a computer science course at the University of Mannheim, Germany, by Cohausz (2022).

Data containing demographic features of students from the 2nd grade of high school (grade 11) in the eastern province high schools of Saudi Arabia were used by Abdalkareem and Min-Allah (2024). It includes demographic features of students, historical academic grades from the first year of high school (grade 10), and historical academic grades, attendance, and behavior for the same sample when they were in the ninth grade. The class label demonstrates the chosen academic pathway for each learner.

Academic records were also used in the context of a bachelor's degree in biology by Lee (2023). The dataset includes demographic information for 569 students majoring in biology (age, gender, and ethnicity), their ACT/SAT scores, the academic plan submitted in the first semester (Fall 2015), the academic plan completed at graduation, and letter grades in freshman-year courses. Only 21% of the students were able to earn a bachelor's degree in biology in four years.

Studying models aimed at classifying student adaptability levels, Nnadi et al. (2024) used a dataset consisting of 1205 rows and 14 features: Gender, Age, Education Level, Institution Type, IT Student, Location, Load-shedding, Financial Condition, Internet Type, Network Type, Class Duration, Self Lms, Device, and the Adaptivity Level. The last is the target and represents the adaptivity level (Low, Moderate, or High) of the student.

Hooper et al. (2023) used multiple choice questions and recording grades or exams, based on three simulated datasets, each one containing 400 simulated student records. The study aims to identify which records in the simulated veterinary student data will predict whether a student will pass or fail a specific course.

The work of Novillo Rangone et al. (2022) enables users to select different types of data as input, including the option to use a survey. Some studies are related to exams, such as the research by Suaza-Medina et al. (2024), which applied tests at various stages of school to determine students' skills gained during the training process, and by Ghimire et al. (2024), who used tasks in general science as predictors for final examination grades.

Alvarez-Garcia et al. (2024) used data collected through three PISA instruments. PISA was also used by Huang et al. (2024) to assess mathematical literacy, as well as three other datasets, including the ASSISTments dataset, which is a comprehensive collection of student interaction data from the ASSISTments online tutoring platform.

Gama Neto et al. (2021) used educational records from the São Paulo State Department of Education (SEDUC-SP), including standardized test scores from SARESP in Portuguese and Mathematics, along with contextual school data such as management history, teacher qualifications, and municipal indicators. Meanwhile, Gusmão et al. (2021) relied on the Microdata from the 2017 Basic Education Census to predict whether schools in Sergipe would meet their projected IDEB targets, considering variables related to school infrastructure, teacher staffing, student–teacher ratios, and employment arrangements.

The study by Aytekin and Saygín (2025) used data related to prerequisites between concepts from three benchmark datasets, including data from Computer Science, Mathematics, and Psychology.

Research using questionnaires was also found. Parkavi et al. (2024) did a significant real-time analysis by collecting the responses from the student's community with a self-evaluated questionnaire, categorized into different sections, including 22 attributes of the knowledge domain, 21 corresponding to the skill domain, and 23 focusing on the attitude domain.

Luo and Wang (2024) uses student questionnaire data from the 2014-2015 China Education Panel Survey (CEPS), which was conducted by Renmin University of China, focusing on 10,279 junior high students, who were initially sampled during the baseline survey, including their life details, growth experiences, and hobbies.

Before closing this section, we present a few examples of practical implications related to educational data, without intending to be exhaustive, and with a focus on recent work. These are small examples that can be strengths or limitations.

- Suaza-Medina et al. (2024): The proposed methodology applies to the educational system of any country as long as it considers the structure of the standardised test and the database to determine the variables that influence academic performance (*strength*);
- Livieris et al. (2023): A limitation of this work is that the proposed framework was evaluated only on two real-world datasets, containing a limited number of attributes, with the scope of predicting the students' performance on the examinations (*limitation*);
- Cavus and Kuzilek (2024): Investigates the relationship between data balancing techniques and counterfactual explanations. RQ1: What is the most appropriate method for generating the counterfactual explanations after balancing? RQ2: How do balancing techniques affect the counterfactual explanations of student success prediction models? (*strength*);
- Choi et al. (2024): the OULAD dataset is specific to the online learning environment, which may limit its generalizability to other contexts (*limitation*);
- Arévalo-Cordovilla and Peña (2024): Limitation of the dataset. The data were collected from a single institution and focused on a specific course, which may have limited the generalization of the findings (*limitation*);
- Aytekin and Saygín (2025): Availability: The training data used for fine-tuning LLMs (Large Language Models) has been made available to other researchers (*strength*);

- Parkavi et al. (2024): tiny sample size of participants makes it difficult to fully explain a few aspects (*limitation*);
- Ghimire et al. (2024), Lee (2023), Q. Liu and Khalil (2024), and Singelmann et al. (2020): Limitation of the dataset (*limitation*);
- Huang et al. (2024): adaptation to different educational contexts. The MFNCD model demonstrated superior performance in four diverse datasets, suggesting good adaptability (*strength*);
- Alvarez-Garcia et al. (2024): Potential for scalability (*strength*);
- Hooper et al. (2023): simulated data with few attributes (*limitation*);
- Gama Neto et al. (2021): combines different data sources to obtain predictor attributes (*strength*).

4.4 Methods

This topic is related to the RQ4 question "What were the methods used to obtain the models in the context of the EDM?". Naturally, the papers present different approaches.

Some works propose to explore an EDM problem by applying methods already known in the literature. However, others present their methods. Among the works that present their methods, some compare the results with existing algorithms in the literature. Likewise, not all algorithms or methods used were evaluated for interpretability. In some cases, the studies make it clear that they considered interpretability in their proposed method or some specific method.

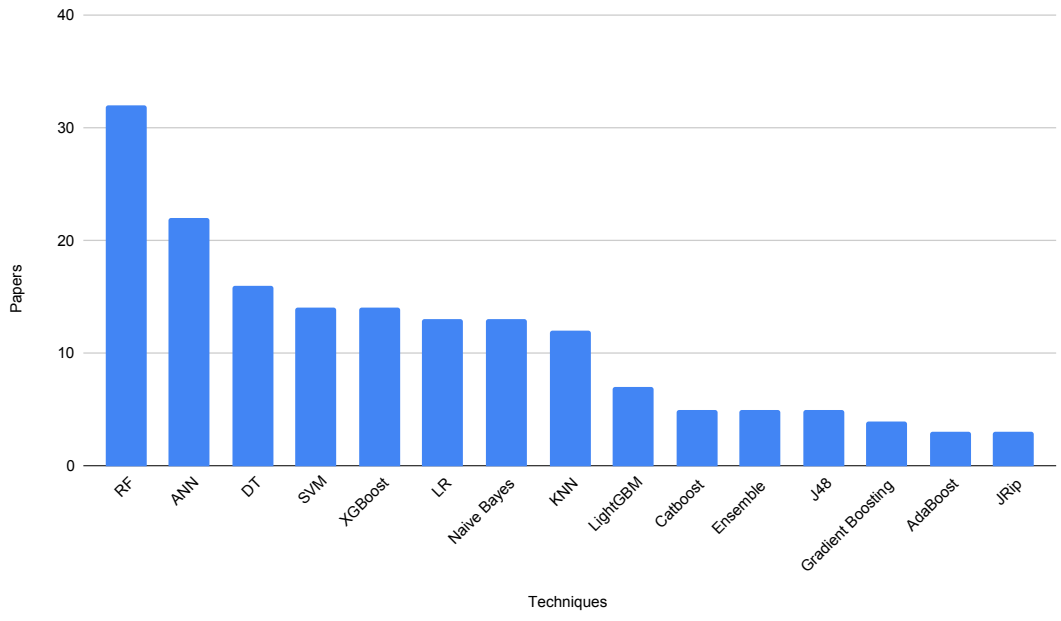

Figure 1: Techniques used to obtain the models at EDM.

Figure 1 presents the algorithms and techniques most frequently used in obtaining DEM models. It is noticed that Random Forest (RF), Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector Machines (SVM), XGBoost (XGB), Logistic Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (KNN) are among the most used techniques. In the following, we present a few examples of practical implications related to ML methods used to obtain models, without intending to be exhaustive or prioritize recent work. These are small examples that can be strengths or limitations.

- Lu et al. (2024): The proposed model is technically complex, but interpretability is a focus to allow targeted interventions. The code, models, and predictions are available under an OSF repository (*strength*);
- Gámez-Granados et al. (2023): Adaptation to different educational contexts: the model was tested in seven different distance higher education courses, suggesting potential for adaptation within this educational level and modality (*strength*);
- Cavus and Kuzilek (2024): Investigates how the model generation process influences the counterfactual explanations generated (*strength*);
- Arévalo-Cordovilla and Peña (2024): Uses feature importance, which is not available for all models (*limitation*);
- Aytekin and Saygín (2025): The only study that used LLM in the context of this mapping (*strength*);
- Thuy2024330: Limited to a specific type of machine learning model (artificial neural networks) (*limitation*);
- Ghimire et al. (2024): Presents a discussion about model limitations (*strength*);
- Hooper et al. (2023): The results show how decisions made by veterinary educators during ML model creation may impact which type of records are shown to be most important (*strength*).

4.5 Strategy to analyze the interpretability

This topic is related to the RQ5 question "Which strategy (method or framework) was used to analyze the interpretability of the models?". As detailed in Section 1, the interpretability of models can be classified by different criteria. In this sense, the selected articles will be presented below, grouped by similarity of the strategies used to explore interpretability. Some articles may be mentioned more than once because they use multiple interpretability methods.

Table 8 presents the main interpretability methods mentioned in the selected studies, prioritizing those that appear in two or more studies. Agnostic and local scope interpretability methods were applied in several of the selected works. We can see there is a prevalence in using SHAP and LIME. Other approaches, such as Intrinsic interpretability, Feature importance / Permutation, ALE, PDP, and Counterfactual explanations, are also related. In a complementary way, Table 9 shows the methods or strategies that have been mentioned only once.

Table 8: List of papers according to the method used for interpretability. The ID is based on Table 4.

Method	Papers IDs	Total
SHAP	3, 5, 8, 10, 12, 14, 15, 17, 19, 20, 21, 26, 30, 32, 33, 49, 52, 56, 58, 61, 62, and 65	22
LIME	6, 8, 12, 13, 15, 16, 26, 27, 30, 32, 40, 46, 53, 56, 60, and 63	16
Intrinsic interpretability	2, 10, 15, 16, 18, 29, 34, 44, 48, 51, 54, and 57	12
Feature importance / Permutation	3, 17, 23, 33, 39, 44, and 64	7
Counterfactual explanations	1, 11, 56, 60, and 65	5
ALE	3, 17, 30, and 56	4
PDP	10, 15, 17, and 30	4
Function Interaction	10 and 17	2
Fuzzy	9 and 36	2
Model grafting	15 and 16	2

Relating Tables 8 and 9 to the level of education and to the educational data, we obtain Table 10. This table allows us to observe the primary applications of interpretability methods across different levels of education and types of educational data. For visualization purposes, we included columns for the two most frequently used methods (SHAP and LIME) and one additional column grouping the remaining methods. Among studies using SHAP, undergraduate courses account for 50%, while basic education represents approximately 41%. Regarding educational data, academic records account for 32%, virtual environments combined with academic records represent 23%, answering exams account for 18%, and virtual environments alone represent 14%. For studies using LIME, undergraduate courses represent approximately 69%, and basic education accounts for about 19%, with the remaining studies distributed across other educational levels. In terms of educational data, virtual environments account for 31%, academic records also represent 31%, and the combination of virtual environments and academic records accounts for 25%.

Additionally, Table 11 is obtained by relating Table 5 to Tables 8 and 9. This table allows the analysis of the distribution of interpretability methods across different interpretability contexts.

Beyond SHAP and LIME, several studies employed additional interpretability strategies. Among global explanation techniques, the Partial Dependence Plot (PDP) (Chou, 2023; Q. Liu & Khalil, 2024; Novillo Rangone et al., 2022; Rangone et al., 2022), the Accumulated Local Effects (ALE) plot (Q. Liu & Khalil, 2024; Nnadi et al., 2024; Novillo Rangone et al., 2022; R. L. C. Silva Filho et al., 2023), and function interaction analysis (Novillo Rangone et al., 2022; Rangone et al., 2022) were identified. Regarding local and model-agnostic approaches, counterfactual explanations were adopted in several studies (Cavus & Kuzilek, 2024; Nnadi et al., 2024; Tsiakmaki & Ragos, 2021). Additionally, Cohausz (2022) proposed an extension of LIME, referred to as Minimally Counterfactual LIME (MC-LIME), aimed at improving the generation of counterfactual explanations.

The other works vary in their characteristics. Alwarthan et al. (2022) used a global interpretability method known as a surrogate global model, and Novillo Rangone et al. (2022) used local surrogate models. The ExpliClas tool, which allows local and global interpretability for some specific models in WEKA, was used in Alonso and Casalino (2019).

Table 9: List of papers for other methods. The ID is based on Table 4.

Method	Papers IDs
Bloom's Taxonomy	4
Attention weights and keys from the Transformer Encoder	7
Global surrogate model	12
Local surrogate	17
Grouped permutation importance (GPI)	19
Attention map (AM)	19
FUMA rules and interviews	22
Teach LLMs to provide explanatory output	24
Word Mover's Distance (WMD)	25
Word Rotator' Similarity (WRS)	25
ELI5	27
Grey Wolf Optimization (GWO)	27
SHAPASH	27
Uncertainty estimation (local and model-specific XAI method)	28
ExpliClas	31
Feature importance based on regularization strength (ANN)	35
Partial least squares structural equation modeling (PLS-SEM)	37
Visualizing how knowledge state changes over time	38
Break-down	39
Degree of Agreement (DOA) analysis	41
Cognitive feature correlation	41
Layer-wise relevance propagation (LRP) into KT	42
Item Response Theory (IRT) model integration	43
Collaborative filtering (CF)	45
Garson's algorithm	46
Lek's profile method	46
HMM model with post-model analysis	47
Box plots and kernel density function plots.	50
Generalized structured component analysis (GSCA)	55
Anchors	56
Click2State and Latent Dirichlet Allocation	59
Gini importance (or mean decrease Gini)	62

Specific interpretability methods for neural networks, such as Garson's algorithm (Garson, 1991) and Lek's profiling method (Lek et al., 1996), were used by the study by Matetic (2019). In the context of artificial neural networks, Luo and Wang (2024) utilized feature importance based on regularization strength, and Lu et al. (2022) employed the LRP method to achieve local and post-hoc interpretability in DLKT models. Jeon et al. (2019) maximized the interpretability of results, limiting the number of topics in a Recurrent Neural Network (RNN), using Click2State and Latent Dirichlet Allocation - LDA. Based on a graph neural network (GNN), the study by Nakagawa et al. (2021) considers interpretability by visualizing how the model predicts changes in student knowledge over time and evaluating the interpretability of its predictions.

In some studies, interpretability was explored, considering specific characteristics of the models. In studies Alharbi (2022) and Al-Jallad et al. (2019), models based on rules were em-

Table 10: Relationship between educational data, level of education, and interpretability methods (SHAP, LIME, and Others).

Educational Data	Level	SHAP	LIME	Others
Virtual environments	Basic Education	-	-	7, 22, 34, 43, 45, 48
	Graduate courses	30	30	30
	Others	-	-	4
	Undergraduate and Graduate	52	-	59
	Undergraduate courses	15	6, 15, 16, 46	1, 9, 11, 16, 28, 31, 37, 46, 47, 51, 54, 55
Academic records	Basic Education	3, 33, 56	56	3, 33, 56
	Undergraduate and Graduate	-	53	-
	Undergraduate courses	10, 19, 49, 62	27, 60, 63	10, 18, 23, 27, 39, 57, 60, 62
Virtual env. + Academic records	Basic Education	8, 21	8, 40	-
	Undergraduate courses	12, 14, 20	12, 13	2, 23, 44
Answering exams	Basic Education	5, 61, 65	-	36, 50, 64, 65
	Others	32	32	25, 38, 42
Questionnaires	Basic Education	-	-	35, 41
	Undergraduate courses	26	26	-
Dynamic (including survey)	Undergraduate and Graduate	17	-	17
Household data	Basic Education	58	-	-
Placement data	All levels	-	-	29
Prerequisites	Undergraduate courses	-	-	24

ployed, whereas in Van Petegem et al. (2023), the specific characteristics of the logistic regression model were analyzed. Another example is the study by Gupta et al. (2022), which considers the characteristics of HMM models.

The work by J. Z. Wang et al. (2017) improved the interpretability of their latent factor model using Bloom's taxonomy (Armstrong, 2010; Krathwohl, 2002). Also, in the context of latent factor, F. Chen and Cui (2020) sought interpretability using a model based on collaborative filtering. Two studies assessed the understanding of their models using explanatory analysis. In Lemay and Doleck (2020), the PLS-SEM approach (Henseler et al., 2016) was applied, while in the study Lemay and Doleck (2022), a generalized analysis of GSCA structured components (Hwang & Takane, 2004) was performed.

The study by Lallé et al. (2021) used a different strategy. Interviews were conducted with experienced instructors to assess the interpretability of a set of rules generated by the FUMA framework. The research by Chau and Phung (2021) proposed a prediction method based on the KNN algorithm. In this case, the interpretability inferred from the KNN is extended to your prediction model. R. L. C. Silva Filho et al. (2023) proposed metrics based on the ALE method and compared them with Permutation Feature Importance (PFI) and SHAP.

Chou (2021) proposes a method of its own to assess interpretability. His method is based on a technique (Model Grafting) that combines DT, DNN, and LIME, from which both global and local interpretations are obtained. The approach enables users to select a desired trade-off between model accuracy and interpretability. Building on Chou (2021), the study Chou (2023) developed a trustworthy and trusted AI framework. The framework was developed and implemented with three main modules: explainable artificial intelligence, safeguard and auditing, and adversarial training. PDP, LIME, and SHAP were also used to provide additional supplementary interpretation for the deep-learning-based DNN model.

Pu et al. (2022) used Word Mover's Distance (WMD) and Word Rotator's Similarity (WRS) to compare the similarity of the knowledge state between ground-truth obtained from different models. The approach utilized by Q. Liu et al. (2018) was visualizing the cognitive diagnosis results in its fuzzy cognitive diagnosis framework (FuzzyCDF). Shen et al. (2023) conducted

Table 11: Relationship between interpretability context and interpretability methods. The IDs are based on Table 4.

Interpretability Context	Method	Paper IDs
Performance Prediction	SHAP	3, 5, 8, 10, 12, 14, 15, 19, 20, 21, 26, 30, 32, 52, 58, 62, 65
	LIME	6, 8, 12, 13, 15, 16, 26, 27, 30, 32, 40, 46, 53, 60, 63
	Intrinsic interpretability	2, 10, 15, 16, 18, 44, 51, 54, 57
	Feature importance / Permutation	3, 23, 39, 44, 64
	Counterfactual explanations	1, 11, 60, 65
	ALE	3, 30
	PDP	10, 15, 30
	Function Interaction	10
	Fuzzy	9
	Model grafting	15, 16
	Global surrogate model	12
	Grouped permutation importance (GPI)	19
	Attention map (AM)	19
	FUMA rules and interviews	22
	ELI5	27
	Grey Wolf Optimization (GWO)	27
	SHAPASH	27
	Uncertainty estimation	28
	ExpliClas	31
	Feature importance (ANN regularization)	35
	PLS-SEM	37
	Break-down	39
	Collaborative filtering (CF)	45
	Garson's algorithm	46
	Lek's profile method	46
	HMM with post-model analysis	47
	GSCA	55
	Anchors	56
	Click2State and LDA	59
	Gini importance	62
Knowledge Tracing	Attention weights (Transformer)	7
	Word Mover's Distance (WMD)	25
	Word Rotator's Similarity (WRS)	25
	Intrinsic interpretability	34, 48
	Visualization of knowledge state changes	38
	LRP	42
	IRT model integration	43
Cognitive Modeling	Fuzzy	36
	Degree of Agreement (DOA) analysis	41
	Cognitive feature correlation	41
Academic Pathways	SHAP	33
	Feature importance / Permutation	33
Adaptability prediction	SHAP	56
	LIME	56
	Counterfactual explanations	56
	ALE	56
	Anchors	56
Career counseling	Intrinsic interpretability	29
Discovery of student profiles	SHAP	61
Instructor Preference	Bloom's Taxonomy	4
Latent Skill	Box plots and kernel density plots	50
Open problem	SHAP	17
	Feature importance / Permutation	17
	ALE	17
	PDP	17
	Function Interaction	17
	Local surrogate	17
Prerequisites	Teach LLMs to provide explanatory output	24
Student loan default	SHAP	49

extensive experiments to show the interpretability of both Learning Process-consistent Knowledge Tracing (LPKT) and LPKT-S. Geng et al. (2023) assessed the interpretability of the enhanced deep cognitive diagnosis model (EDCDM) by showing the distribution of students' skill states through box plots and kernel density function plots.

In the following, we present a few examples of practical implications related to interpretability methods, without intending to be exhaustive, and with a focus on recent work. These are small examples that can be strengths or limitations.

- Colpo et al. (2024): Analyzes interpretability for different moments (before and during the COVID-19 pandemic) (*strength*);
- Zanellati et al. (2024): Compares different post hoc explainability techniques (*strength*);
- Hooper et al. (2023): The discussion focuses on how the different approaches to handling missing data and the different interpretability techniques (Gini versus SHAP) provide distinct insights into the importance of features for the Random Forest model (*strength*);
- Arévalo-Cordovilla and Peña (2024): Only considers interpretability using feature importance, which is not available for all models. Does not perform further analysis or comparisons (*limitation*);
- Parkavi et al. (2024): Does not delve into the analysis of the results obtained by the interpretability methods (*limitation*);
- Abdalkareem and Min-Allah (2024): The SHAP values were calculated only for the Random Forest model in this study. Application of interpretability methods in a very limited way (*limitation*);
- Lee (2023): Applied only to one of the models (Random Forest) (*limitation*);
- Alvarez-Garcia et al. (2024): Interpretability applied only to part of the experiments (*limitation*);
- Kumar and Sharma (2020): The study only applies the LIME method, without much detail, to two samples. One sample with satisfactory performance and another with unsatisfactory performance. A very brief analysis of the two main features involved is made (*limitation*).
- C. Carvalho et al. (2023): It goes beyond the simple use of explanatory methods, seeking to analyze a set of explanations with the aim of comparing prediction models in terms of interpretability (*strength*).

4.6 Metrics to assess interpretability

This topic is related to the RQ6 question "What metrics were used to assess the interpretability of a model?". The proposal is to identify studies that have initiatives to measure the interpretability of models, allowing us to differentiate between models and determine which one is more interpretable than another. Few studies, however, presented these characteristics. Interpretability, for the most part, was considered only as a way of explaining the model or specific predictions.

Few studies have attempted to measure the interpretability of machine learning models. The work of Alharbi (2022) proposed measuring interpretability using a complexity metric. The study mentions that in rule-based models, some metrics have been proposed to measure the interpretability of models, for example, in Garcia et al. (2009) and Nauck (2003). In this sense, to increase the interpretability of the model, Alharbi (2022) limited the number of rules per class, justified by the fact that humans can deal with a maximum of 7±2 cognitive entities at the same time (Miller, 1956). The results demonstrate the ability to obtain models with high prediction accuracy and interpretability, compared to black box models.

Al-Jallad et al. (2019) used interpretable methods such as J48, BFTree, LADTree, RandomTree, Simple-CART, REP-tree, JRip, Ridor, and PART. As part of the evaluation, it proposed an interpretability measure based on the size and complexity of the rules. For each experiment, the interpretability measurements of each model are presented, along with an analysis of the effects of the preprocessing step on these measurements. The results demonstrate that the J48 algorithm stood out for its ability to apply an appropriate trade-off between accuracy and interpretability.

The study by C. Carvalho et al. (2023) proposes an interpretability metric called agreement percentage. The proposal involves analyzing a set of LIME explanations using unsupervised ML to find the core explanations that best describe the behavior of the original model. Subsequently, predictions from the model are compared with predictions based on the core explanations, allowing for the identification of regions where the model's response aligns with the core explanations that represent the same model. This approach enables the differentiation of models based on aspects of interpretability.

The work by Nakagawa et al. (2021) is based on the Graph Neural Network (GNN) technique. It proposed to evaluate interpretability based on two points directly related to the model update. The first is to evaluate whether the model updates only the concept related to the answered concept in each time interval. The second is whether the update is reasonable with the given graph structure. In this work, a brief analysis of these aspects is presented; simultaneously, it aims to justify the better interpretability of the GNN model for others.

Following another strategy, the work of Lallé et al. (2021) proposed evaluating interpretability through instructor interviews. The rules mined by the FUMA framework were presented to instructors, who were asked to interpret them based on their experience and expertise. Results pointed to the high interpretability of the rules generated by FUMA. In particular, the interview results showed substantial agreement between FUMA and experienced instructors, and the instructors provided ideas for adaptive feedback on low-performance rules.

Chou (2021, 2023) proposed assessing interpretability based on a grafting technique, combining ML and interpretability methods. The approach allows users to select the desired trade-off between accuracy and interpretability for the model. Gámez-Granados et al. (2023) assesses the interpretability of FlexNSLVOrd in a qualitative way, comparing it to other rule-based models. The main aspects considered were the simplicity and representativeness of the rules, as well as the average number of generated rules, where a few rules are generally associated with greater interpretability. Cavus and Kuzilek (2024) have assessed essential properties for counterfactual explanations, such as sparsity, minimality, validity, proximity, and plausibility.

Aytekin and Saygín (2025) used their own domain knowledge to manually create explanations and compare them to the explanations of Fine-tuned LLM, with the evaluation based on

the validity and similarity of the explanations. The LLM explanation information was checked, and a similarity score was generated. The interpretability assessment by F. Liu et al. (2024) was mainly qualitative, based on the presentation and explanation of fuzzy rules, hidden semantics, and visualizations of results. The comparison of interpretability between models was conducted descriptively, highlighting the lack of intrinsic interpretability in most deep learning-based KT models and presenting FDKT as an approach that explicitly seeks to improve this aspect.

In summary, this section highlighted eleven studies that have presented initiatives to measure the interpretability of models, representing 17% of the selected articles. As can be seen in Table 12 five studies do so, focusing on rules (Alharbi, 2022; Al-Jallad et al., 2019; Gámez-Granados et al., 2023), including interviews (Lallé et al., 2021) and fuzzy rules (F. Liu et al., 2024). A specific work considers models based on Graph Neural Network (GNN) (Nakagawa et al., 2021).

Four studies address the use of agnostic interpretability methods, using either the grafting technique (Chou, 2021, 2023), counterfactual explanations (Cavus & Kuzilek, 2024), or clusters of LIME explanations (C. Carvalho et al., 2023). Finally, a study using LLMs assessed model interpretability using domain knowledge to evaluate the validity and similarity of the explanations (Aytekin & Saygín, 2025).

Table 12: Relationship between metrics to assess interpretability, interpretability context, educational data, and level of education. The ID is based on Table 4.

Measure of interpretability	Interpretability context	Educational data	Level of education	Paper ID
Complexity of rules	Academic records	Performance Prediction	Undergraduate courses	18
Complexity of rules	Academic records	Performance Prediction	Undergraduate courses	57
Simplicity and quantity of rules	Virtual environments	Performance Prediction	Undergraduate courses	9
Interpreting rules through interviews with instructors	Virtual environments	Performance Prediction	Basic Education	22
Qualitative analysis of fuzzy rules	Virtual environments	Performance Prediction	Graduate courses	30
GNN model update	Answering exams	Knowledge Tracing	Others	38
Grafting Technique (accuracy–interpretability trade-off selection)	Virtual environments	Performance Prediction	Undergraduate courses	16
Grafting Technique (accuracy–interpretability trade-off selection)	Virtual environments	Performance Prediction	Undergraduate courses	15
Properties for counterfactual explanations	Virtual environments	Performance Prediction	Undergraduate courses	11
Agreement percentage - Clusters of LIME explanations	Academic records	Performance Prediction	Undergraduate courses	63
Validity and similarity using domain knowledge and LLMs	Prerequisites	Prerequisites	Undergraduate courses	24

4.7 Objective when addressing interpretability

This topic is related to the RQ7 question "What is the purpose or objective when addressing interpretability?". The proposal is to identify the reasons or objectives behind the use of interpretability. Are the authors aiming to understand the model's decisions to enhance transparency? Improve stakeholder trust? Analyze or mitigate unfairness?

Categorizing these objectives can be highly subjective and dependent on author declarations, involving different classes of stakeholders and desiderata such as acceptance, effectiveness, fairness, transparency, trust, legal compliance, and many others (Langer et al., 2021). For this research question, we will focus on fairness, as recent research has linked it to interpretability (Deck et al., 2024). In the following, we present general aspects of some studies to exemplify objectives when addressing interpretability. This list is not exhaustive, representing only a part of the selected works. Later, a highlight will be made on cases related to algorithmic fairness.

First, we can mention some cases in summary form, with a simple objective of improving model transparency (Abdalkareem & Min-Allah, 2024; Choi et al., 2024; Colak Oz et al., 2023; Huang et al., 2024; Islam et al., 2024; Luo & Wang, 2024; Parkavi et al., 2024).

In other cases, we can highlight the intention to identify the reasons for the detection of prerequisites (Aytekin & Saygín, 2025), understand the impact of each variable on the model decision (Arévalo-Cordovilla & Peña, 2024; Nnadi et al., 2024), identify key dropout patterns and how these patterns compared before and during the pandemic (Colpo et al., 2024), clusters explicability (Alvarez-Garcia et al., 2024), assess the concordance between model decisions and explanations (C. Carvalho et al., 2023), and other related to the influence of predictors (Lee, 2023; Singelmann et al., 2020; Suaza-Medina et al., 2024; Y. Wang et al., 2023).

There are also examples where the author demonstrates attention with respect to stakeholders. Lu et al. (2024) makes the model more accessible to researchers and educators, enabling them to enhance pedagogical practices and support for students in online learning environments. Gámez-Granados et al. (2023) utilized interpretability with the primary objective of providing understandable results that can be easily understood by teachers, whereas Cavus and Kuzilek (2024) seeks to establish trust from students and teachers, aiming for features such as stability and robustness. Other studies also fit into these cases, such as Ghimire et al. (2024), F. Liu et al. (2024), F. Liu et al. (2023), and Q. Liu and Khalil (2024), including facilitating communication with stakeholders (Hooper et al., 2023), issuing alerts (Qu et al., 2022), and achieving interpretability as a basis for constructing personalized interventions (Cohausz, 2022).

Specifically regarding fairness, there are a few cases that can be highlighted due to their relationship with interpretability:

- Livieris et al. (2023): mention fairness and bias. The adoption of SHAP ensures that the importance scores are fair and unbiased, while the adoption of LIME provides a flexible, fast, and reliable technique for interpreting individual predictions.
- Zanellati et al. (2024): Interpretability techniques are applied to increase confidence, and a fairness analysis was conducted to make models more effective and equitable. The study expands the use of deep learning in educational data mining, achieving high accuracy while improving model transparency and fairness;
- Tsiakmaki and Ragos (2021): The main objective is to demonstrate the practical usefulness of understandable learning models in educational environments, and a deep machine learning model with a fair predictive performance was created;

In the following, we present a few examples of practical implications related to interpretability objectives, without intending to be exhaustive, and with a focus on recent work. These are small examples that can be strengths or limitations:

- Zanellati et al. (2024): In addition to predictive performance, it emphasizes the ethical integrity of models, especially fairness (*strength*);
- Thuy and Benoit (2024): The framework incorporates the involvement of human experts in the decision process through classification with rejection for uncertain observations (*strength*);
- Q. Liu and Khalil (2024): The study does not investigate the reception of XAI feedback by stakeholders (*limitation*);
- Y. Wang et al. (2023): It includes policy recommendations to help university administrators and bank staff intervene early to reduce the risk of default by university students (*strength*)

5 Discussion and future directions

This Section discusses the findings and potential future work, with a focus on the interpretability field. The main interest of this mapping was the aspect of model interpretability in the EDM context. The selected studies are presented in various contexts, including the problems addressed, education levels, data types, machine learning techniques, objectives, and strategies employed to address interpretability. Despite the diversity of contexts and approaches, the analysis reveals structural patterns that allow for a broader reflection on how interpretability has been positioned within EDM research.

Given the increasing prominence of explainable artificial intelligence (XAI) in recent years (Adadi & Berrada, 2018; Linardatos et al., 2021; ŞAHiN et al., 2025), one might expect interpretability to be a primary design objective in EDM studies. However, the results of this mapping suggest that interpretability is frequently treated as a complementary analysis step rather than as a core methodological concern. This finding is consistent with broader observations that interpretability techniques are still rarely integrated as a central component of machine learning workflows (Linardatos et al., 2021). The predominance of performance prediction studies and the widespread use of post hoc methods indicate that model transparency is often pursued after model development rather than integrated into model design.

The findings also reveal structural relationships between EDM contexts and interpretability choices. Knowledge Tracing studies, for example, tend to rely on model-specific explanations such as attention mechanisms or IRT-based analyses, whereas Performance Prediction studies predominantly employ model-agnostic techniques. This suggests that interpretability strategies are frequently influenced by model architecture rather than by stakeholder needs or educational objectives.

Consistent with the previously identified gap, very few studies propose evaluating interpretability through formal metrics. This gap corroborates previous findings that highlight the absence or difficulty of defining appropriate metrics for evaluating interpretability methods (Alamri & Alharbi, 2021; D. V. Carvalho et al., 2019; Vieira & Digiampietri, 2022), raising concerns about how interpretability claims are substantiated in EDM research.

These structural patterns are reflected in the operational practices observed in the analyzed studies. Typically, reports are limited to using an interpretability method. Some research focuses on prediction performance and presents a closing Section showing that it is possible to interpret the obtained models. In other cases, interpretability is mentioned as part of the work, but only some details are presented. Therefore, from this aspect, interpretability is not the central objective in many studies.

In cases where the study considers several ML techniques, it is not always clear whether interpretability has been applied to all of them. No analysis compares the interpretability between models obtained for different ML techniques. From this perspective, future research should prioritize the comparison and differentiation of models based on their interpretability.

Very few studies propose evaluating interpretability from the perspective of an interpretability metric. Some cases involving this approach were for intrinsically interpretable models or those that required human intervention, and only one case used an agnostic method. Exploring interpretability metrics is considered relevant to support the decision to choose models.

Fairness is an important stakeholder concern in terms of interpretability; however, very few studies explore the relationship between fairness and interpretability. Additionally, stakeholders must be involved in the process.

After identifying the limitations described above, we present specific points that should be addressed in future work on interpretability in EDM. Whenever possible, studies from this mapping are cited to illustrate initiatives that align with, or show progress toward, these recommendations.

Studies should focus on interpretability in EDM, rather than treating it as a secondary aspect of model evaluation. It can be addressed, for example, by identifying patterns and trade-offs (Chou, 2021, 2023), by comparing prediction models in terms of interpretability (C. Carvalho et al., 2023), by analyzing the interpretability for different moments (before and during the COVID-19 pandemic) (Colpo et al., 2024), or by comparing different post hoc explainability techniques (Zanellati et al., 2024).

Studies should develop and apply interpretability metrics that enable the quantification of differences between models, thereby providing a stronger basis for model selection. This can be addressed by analyzing the complexity of rules in intrinsically interpretable models (Alharbi, 2022; Al-Jallad et al., 2019), but it should prioritize the quantification of interpretability for black-box models based on agnostic methods (C. Carvalho et al., 2023; Chou, 2023). Domain knowledge can also be used to evaluate the validity and similarity of the explanations (Aytekin & Saygín, 2025), as well as interviews with experienced instructors to assess interpretability (Lallé et al., 2021).

Studies should investigate how different stages of the EDM pipeline impact interpretability, ensuring that methodological choices do not compromise model transparency. This can be achieved, for example, by examining the relationship between balancing techniques and explanations (Cavus & Kuzilek, 2024), or the relationship between various approaches to handling missing data and different interpretability techniques (Hooper et al., 2023).

Studies should address the relationship between fairness and interpretability. Although initiatives can be found in some way (Livieris et al., 2023; Tsiakmaki & Ragos, 2021; Zanellati et al., 2024), it is necessary to go further, addressing fairness analysis, reporting, and mitigation in EDM applications (Deck et al., 2024).

Studies should ensure that EDM application projects actively involve stakeholders throughout the development process to align interpretability goals with practical needs, including correctly understanding model explanations, and encouraging the creation of usage guides.

Ultimately, studies should investigate the implications of deploying EDM models in production, particularly how future updates may impact interpretability and fairness.

6 Final considerations

This paper conducted a Systematic Literature Mapping (SLM) on model interpretability within the context of Educational Data Mining (EDM). A research protocol was defined in accordance with the guidelines of Kitchenham and Charters (2007), resulting in 65 selected studies.

The study provides a structured overview of how interpretability has been addressed in EDM, identifying methodological trends, practical implications, and research gaps. By mapping interpretability methods, their applications, and their alignment with different educational contexts, this work offers insights that can guide both researchers and practitioners. However, some limitations must be acknowledged.

The primary limitation of this mapping is that it focuses on studies that explicitly identify themselves as part of the EDM field. Consequently, studies that employ interpretability methods in educational contexts but do not fall within the EDM scope may not have been included.

By addressing the aspects discussed in the previous section, future research can advance the field toward a more systematic and impactful understanding of interpretability in EDM. Such efforts can foster transparency, trust, and fairness, support more informed decision-making by educational institutions, enhance student support systems, and promote accountability in AI-driven educational applications.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Abdalkareem, M., & Min-Allah, N. (2024). Explainable models for predicting academic pathways for high school students in saudi arabia. *IEEE Access*, *12*, 30604–30626. https://doi.org/10.1109/ACCESS.2024.3369586 [GS Search].

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 [GS Search].

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, *9*, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368 [GS Search].

Alharbi, B. (2022). Back to basics: An interpretable multi-class grade prediction framework. *Arabian Journal for Science and Engineering*, *47*(2), 2171–2186. https://doi.org/10.1007/s13369-021-06153-x [GS Search].

Al-Jallad, N. T., Ning, X., Khairalla, M. A., & Al-Qaness, M. A. (2019). Rule mining models for predicting dropout/ stopout and switcher at college using satisfaction and ses features. *International Journal of Management in Education*, *13*(2), 97–118. https://doi.org/10.1504/IJMIE.2019.098182 [GS Search].

Alonso, J. M., & Casalino, G. (2019). Explainable artificial intelligence for human-centric data analysis in virtual learning environments. *Communications in Computer and Information Science*, *1091*, 125–138. https://doi.org/10.1007/978-3-030-31284-8_10 [GS Search].

Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles. an explainable cluster analysis approach to PISA 2022. *Computers & Education*, *223*, 105166. https://doi.org/10.1016/j.compedu.2024.105166 [GS Search].

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, *10*, 107649–107668. https://doi.org/10.1109/ACCESS.2022.3211070 [GS Search].

Araujo, I. (2021). Uma revisão sobre o uso de frameworks de interpretabilidade em aprendizado de máquina. *Anais do XIV Encontro Unificado de Computação do Piauí e XI Simpósio de Sistemas de Informação*, 105–112. https://doi.org/10.5753/enucompi.2021.17760 [GS Search].

Arévalo-Cordovilla, F. E., & Peña, M. (2024). Comparative analysis of machine learning models for predicting student success in online programming courses: A study based on lms data and external factors. *Mathematics*, *12*(20), 3272. https://doi.org/10.3390/math12203272 [GS Search].

Armstrong, P. (2010). Bloom's taxonomy. *Vanderbilt University Center for Teaching*, *12*(05), 2023. [GS Search].

Aytekin, M. C., & Saygín, Y. (2025). Discovering prerequisite relations using large language models. *Interactive Learning Environments*, *33*(2), 1670–1688. https://doi.org/10.1080/10494820.2024.2375338 [GS Search].

Baker, R. S. J. d. (2010). Data mining. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education (third edition)* (Third Edition, pp. 112–118). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-044894-7.01318-X [GS Search].

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657 [GS Search].

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*, 537–553. https://doi.org/10.1007/s10639-017-9616-z [GS Search].

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228 [GS Search].

Carvalho, C., Mattos, J., & Aguiar, M. (2023). Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior. *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, 1191–1201. https://doi.org/10.5753/sbie.2023.234435 [GS Search].

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). https://doi.org/10.3390/electronics8080832 [GS Search].

Cavus, M., & Kuzilek, J. (2024). An effect analysis of the balancing techniques on the counterfactual explanations of student success prediction models. *Journal of Measurement and Evaluation in Education and Psychology*, *15*, 302–317. https://doi.org/10.21031/epod.1526704 [GS Search].

Chau, V. T. N., & Phung, N. H. (2021). A cumulative increasing kemelized nearest-neighbor bagging method for early course-level study performance prediction. *2021 7th International*

Conference on Engineering, Applied Sciences and Technology (ICEAST), 91–96. https://doi.org/10.1109/iceast52143.2021.9426259 [GS Search].

Chen, F., & Cui, Y. (2020). LogCF: Deep collaborative filtering with process data for enhanced learning outcome modeling. *Journal of Educational Data Mining*, *12*(4), 66–99. https://doi.org/10.5281/zenodo.4399685 [GS Search].

Chen, S., Pian, Y., & Zheng, Y. (2023). Challenges and strategies for designing more effective educational data mining applications. *2023 Twelfth International Conference of Educational Innovation through Technology (EITT)*, 175–179. https://doi.org/10.1109/EITT61659.2023.00040 [GS Search].

Choi, W. C., Lam, C.-T., & Mendes, A. J. (2024). Analyzing the interpretability of machine learning prediction on student performance using shapley additive explanations. *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 1–8. https://doi.org/10.1109/TALE62452.2024.10834292 [GS Search].

Chou, T.-N. (2021). Apply explainable AI to sustain the assessment of learning effectiveness. *IMCIC 2021 - 12th International Multi-Conference on Complexity, Informatics and Cybernetics, Proceedings*, *2*, 113–118. [GS Search].

Chou, T.-N. (2023). Apply an integrated responsible AI framework to sustain the assessment of learning effectiveness. *International Conference on Computer Supported Education, CSEDU - Proceedings*, *2*, 142–149. https://doi.org/10.5220/0012058400003470 [GS Search].

Cohausz, L. (2022). Towards real interpretability of student success prediction combining methods of XAI and social science. *Proceedings of the 15th International Conference on Educational Data Mining*, 361–367. https://doi.org/10.5281/zenodo.6853069 [GS Search].

Colak Oz, H., Güven, Ç., & Nápoles, G. (2023). School dropout prediction and feature importance exploration in Malawi using household panel data: Machine learning approach. *Journal of Computational Social Science*, *6*(1), 245–287. https://doi.org/10.1007/s42001-022-00195-3 [GS Search].

Colpo, M. P., Primo, T. T., & Aguiar, M. S. (2024). Lessons learned from the student dropout patterns on covid-19 pandemic: An analysis supported by machine learning. *British Journal of Educational Technology*, *55*(2), 560–585. https://doi.org/10.1111/bjet.13380 [GS Search].

Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A critical survey on fairness benefits of explainable AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1579–1595. https://doi.org/10.1145/3630106.3658990 [GS Search].

Dermeval, D., Coelho, J., & Bittencourt, I. I. (2020). Mapeamento sistemático e revisão sistemática da literatura em informática na educação. In P. A. Jaques, S. Siqueira, I. Bittencourt, & M. Pimentel (Eds.), *Metodologia de pesquisa científica em informática na educação: Abordagem quantitativa* (Vol. 2). SBC. https://ceie.sbc.org.br/metodologia/ [GS Search].

Gama Neto, M., Vasconcelos, G., & Zanchettin, C. (2021). Mineração de dados aplicada à predição do desempenho de escolas e técnicas de interpretabilidade dos modelos. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 773–782. https://doi.org/10.5753/sbie.2021.217421 [GS Search].

Gámez-Granados, J. C., Esteban, A., Rodriguez-Lozano, F. J., & Zafra, A. (2023). An algorithm based on fuzzy ordinal classification to predict students' academic performance. *Applied*

Intelligence, *53*(22), 27537–27559. https://doi.org/10.1007/s10489-023-04810-2 [GS Search].

Garcia, S., Fernandez, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, *13*(10), 959–977. https://doi.org/10.1007/s00500-008-0392-y [GS Search].

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, *6*(4), 46–51. [GS Search].

Geng, J., Yang, H., & Hu, S. (2023). Noise-filtering enhanced deep cognitive diagnosis model for latent skill discovering. *Intelligent Automation and Soft Computing*, *37*(2), 1311–1324. https://doi.org/10.32604/iasc.2023.038481 [GS Search].

Ghimire, S., Abdulla, S., Joseph, L. P., Prasad, S., Murphy, A., Devi, A., Barua, P. D., Deo, R. C., Acharya, R., & Yaseen, Z. M. (2024). Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100331 [GS Search].

Guleria, P., & Sood, M. (2023). Explainable AI and machine learning: Performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies*, *28*(1), 1081–1116. https://doi.org/10.1007/s10639-022-11221-2 [GS Search].

Gupta, A., Garg, D., & Kumar, P. (2022). Mining sequential learning trajectories with hidden markov models for early prediction of at-risk students in e-learning environments. *IEEE Transactions on Learning Technologies*, *15*(6), 783–797. https://doi.org/10.1109/TLT.2022.3197486 [GS Search].

Gusmão, R., Gusmão, C., & Dias, M. (2021). A qualidade da educação para além do IDEB: Um estudo através de técnicas de mineração de dados. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 803–812. https://doi.org/10.5753/sbie.2021.218177 [GS Search].

Hegazi, M. O., & Abugroon, M. A. (2016). The state of the art on educational data mining in higher education. *International Journal of Computer Trends and Technology*, *31*(1), 46–56. [GS Search].

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management & Data Systems*, *116*(1), 2–20. https://doi.org/10.1108/IMDS-09-2015-0382 [GS Search].

Hooper, S. E., Hecker, K. G., & Artemiou, E. (2023). Using machine learning in veterinary medical education: An introduction for veterinary medicine educators. *Veterinary Sciences*, *10*(9). https://doi.org/10.3390/vetsci10090537 [GS Search].

Hooshyar, D., Huang, Y.-M., & Yang, Y. (2022). A three-layered student learning model for prediction of failure risk in online learning. *Human-centric Computing and Information Sciences*, *12*. https://doi.org/10.22967/HCIS.2022.12.028 [GS Search].

Huang, T., Geng, J., Yang, H., Hu, S., Ou, X., Hu, J., & Yang, Z. (2024). Interpretable neuro-cognitive diagnostic approach incorporating multidimensional features. *Knowledge-Based Systems*, *304*. https://doi.org/10.1016/j.knosys.2024.112432 [GS Search].

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, *69*(1), 81–99. https://doi.org/10.1007/BF02295841 [GS Search].

Islam, M. R., Nitu, A. M., Marjan, M. A., Uddin, M. P., Afjal, M. I., & Al Mamun, M. A. (2024). Enhancing tertiary students' programming skills with an explainable educational data mining approach. *PLoS ONE*, *19*(9 September). https://doi.org/10.1371/journal.pone.0307536 [GS Search].

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, *27*(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6 [GS Search].

Jeon, B., Shafran, E., Breitfeller, L., Levin, J., & Rosé, C. P. (2019). Time-series insights into the process of passing or failing online university courses using neural-induced interpretable student states. *Proceedings of the 12th International Conference on Educational Data Mining*, 330–335. https://doi.org/10.48550/arXiv.1905.00422 [GS Search].

Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost – an ensemble machine learning model for prediction and classification of student academic performance. *Advances in Data Science and Adaptive Analysis*, *13*(03N04). https://doi.org/10.1142/S2424922X21410023 [GS Search].

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074 [GS Search].

Kitchenham, B. A., & Charters, S. (2007, July). *Guidelines for performing systematic literature reviews in software engineering* (tech. rep. No. EBSE 2007-001). Keele University and Durham University Joint Report. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf [GS Search].

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2 [GS Search].

Kumar, P., & Sharma, M. (2020). Predicting academic performance of international students using machine learning techniques and human interpretable explanations using lime—case study of an indian university. *Advances in Intelligent Systems and Computing*, *1087*, 289–303. https://doi.org/10.1007/978-981-15-1286-5_25 [GS Search].

Lallé, S., Yalçın, Ö. N., & Conati, C. (2021). Combining data-driven models and expert knowledge for personalized support to foster computational thinking skills. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 375–385. https://doi.org/10.1145/3448139.3448175 [GS Search].

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? – a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473 [GS Search].

Lee, Y. (2023). Identifying prerequisite courses in undergraduate biology using machine learning. *Journal of Data Science*, *21*(4), 745–760. https://doi.org/10.6339/22-JDS1068 [GS Search].

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological modelling*, *90*(1), 39–52. [GS Search].

Lemay, D. J., & Doleck, T. (2020). Grade prediction of weekly assignments in MOOCS: Mining video-viewing behavior. *Education and Information Technologies*, *25*(2), 1333–1342. https://doi.org/10.1007/s10639-019-10022-4 [GS Search].

Lemay, D. J., & Doleck, T. (2022). Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. *Interactive Learning Environments*, *30*(10), 1782–1793. https://doi.org/10.1080/10494820.2020.1746673 [GS Search].

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1). https://doi.org/10.3390/e23010018 [GS Search].

Liu, F., Bu, C., Zhang, H., Wu, L., Yu, K., & Hu, X. (2024). FDKT: Towards an interpretable deep knowledge tracing via fuzzy reasoning. *ACM Transactions on Information Systems*, *42*(5). https://doi.org/10.1145/3656167 [GS Search].

Liu, F., Zhang, H., & Huang, W. (2023). Interpreting learner success: Enhancing knowledge tracing with attention-based IRT models in modern education. *Proceedings of the 13th International Conference on Information Technology in Medicine and Education, ITME 2023*, 649–653. https://doi.org/10.1109/ITME60234.2023.00135 [GS Search].

Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., Chen, Z., & Hu, G. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology*, *9*(4). https://doi.org/10.1145/3168361 [GS Search].

Liu, Q., & Khalil, M. (2024). Explainable AI in learning analytics: Improving predictive models and advancing transparency trust. *2024 IEEE Global Engineering Education Conference (EDUCON)*, 1–7. https://doi.org/10.1109/EDUCON60312.2024.10578733 [GS Search].

Liu, X., Zambrano, A. F., Baker, R. S., Barany, A., Ocumpaugh, J., Zhang, J., Pankiewicz, M., Nasiar, N., & Wei, Z. (2025). Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics*, *12*(1), 169–185. https://doi.org/10.18608/jla.2025.8575 [GS Search].

Livieris, I. E., Karacapilidis, N., Domalis, G., & Tsakalidis, D. (2023). An advanced explainable and interpretable ML-based framework for educational data mining. *Lecture Notes in Networks and Systems*, *769 LNNS*, 87–96. https://doi.org/10.1007/978-3-031-42134-1_9 [GS Search].

Lu, Y., Tong, L., & Cheng, Y. (2024). Advanced knowledge tracing: Incorporating process data and curricula information via an attention-based framework for accuracy and interpretability. *Journal of Educational Data Mining*, *16*(2), 58–84. https://doi.org/10.5281/zenodo.13712553 [GS Search].

Lu, Y., Wang, D., Chen, P., Meng, Q., & Yu, S. (2022). Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00297-z [GS Search].

Luo, Y., & Wang, Z. (2024). Feature mining algorithm for student academic prediction based on interpretable deep neural network. *12th International Conference on Information and Education Technology, ICIET 2024*, 1–5. https://doi.org/10.1109/ICIET60671.2024.10542709 [GS Search].

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access*, *11*, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344 [GS Search].

Matetic, M. (2019). Mining learning management system data using interpretable neural networks. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1282–1287. https://doi.org/10.23919/MIPRO.2019.8757113 [GS Search].

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81. https://doi.org/10.1037/h0043158 [GS Search].

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book [GS Search].

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008 [GS Search].

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2021). Graph-based knowledge tracing: Modeling student proficiency using graph neural networks. *Web Intelligence*, *19*(1-2), 87–102. https://doi.org/10.3233/WEB-210458 [GS Search].

Nauck, D. D. (2003). Measuring interpretability in rule-based classification systems. *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03.*, *1*, 196–201. https://doi.org/10.1109/FUZZ.2003.1209361 [GS Search].

Nnadi, L. C., Watanobe, Y., Rahman, M. M., & John-Otumu, A. M. (2024). Prediction of students' adaptability using explainable AI in educational machine learning models. *Applied Sciences*, *14*(12), 5141. https://doi.org/10.3390/app14125141 [GS Search].

Novillo Rangone, G., Pizarro, C., & Montejano, G. (2022). Automation of an educational data mining model applying interpretable machine learning and auto machine learning. In Á. Rocha, D. Barredo, P. C. López-López, & I. Puentes-Rivera (Eds.), *Communication and smart technologies* (pp. 22–30). Springer Singapore. https://doi.org/10.1007/978-981-16-5792-4_3 [GS Search].

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, *93*, 106384. https://doi.org/10.1016/j.asoc.2020.106384 [GS Search].

Pandian, B. R., Aziz, A. A., Subramaniam, H., & Nawi, H. S. A. (2024). Exploring the role of machine learning in forecasting student performance in education: An in-depth review of literature. *Multidisciplinary Reviews*, *6*, 2023ss043. https://doi.org/10.31893/multirev.2023ss043 [GS Search].

Pantazatos, D., Trilivas, A., Meli, K., Kotsifakos, D., & Douligeris, C. (2024). Machine learning and explainable artificial intelligence in education and training – status and trends. In L. A. Maglaras & C. Douligeris (Eds.), *Wireless internet* (pp. 110–122). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58053-6_8 [GS Search].

Parkavi, R., Karthikeyan, P., & Sheik Abdullah, A. (2024). Enhancing personalized learning with explainable AI: A chaotic particle swarm optimization based decision support system. *Applied Soft Computing*, *156*. https://doi.org/10.1016/j.asoc.2024.111451 [GS Search].

Pei, B., & Xing, W. (2022). An interpretable pipeline for identifying at-risk students. *Journal of Educational Computing Research*, *60*(2), 380–405. https://doi.org/10.1177/07356331211038168 [GS Search].

Peña-Ayala, A. (2014). *Educational data mining: Applications and trends* (Vol. 524). Springer International Publishing. https://doi.org/10.1007/978-3-319-02738-8 [GS Search].

Pu, Y., Wu, W., Peng, T., Liu, F., Liang, Y., Yu, X., Chen, R., & Feng, P. (2022). Embedding cognitive framework with self-attention for interpretable knowledge tracing. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-22539-9 [GS Search].

Qu, Y., Li, F., Li, L., Dou, X., & Wang, H. (2022). Can we predict student performance based on tabular and textual data? *IEEE Access*, *10*, 86008–86019. https://doi.org/10.1109/ACCESS.2022.3198682 [GS Search].

Queiroga, E. M., Santana, D., Silva, M., Aguiar, M., Santos, V., Mello, R. F., Bittencourt, I. I., & Cechinel, C. (2024). Anticipating student abandonment and failure: Predictive models in high school settings. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 351–364). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64302-6_25 [GS Search].

Rachha, A., & Seyam, M. (2023). Explainable AI in education: Current trends, challenges, and opportunities. *SoutheastCon 2023*, 232–239. https://doi.org/10.1109/SoutheastCon51012.2023.10115140 [GS Search].

Raji, N., Kumar, R. M. S., & Biji, C. (2023). Closing the gap: Exploring the untapped potential of machine learning in deaf students and hearing students' academic performance. *International Journal of Advanced Technology and Engineering Exploration*, *10*(108), 1449–1475. https://doi.org/10.19101/IJATEE.2023.10101685 [GS Search].

Rangone, G. N., Montejano, G. A., Garis, A. G., Pizarro, C. A., & Molina, W. R. (2022). An educational data mining model based on auto machine learning and interpretable machine learning. *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*, 1–6. https://doi.org/10.1109/GlobConPT57482.2022.9938243 [GS Search].

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075 [GS Search].

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355 [GS Search].

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*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532 [GS Search].

ŞAHiN, E., Arslan, N. N., & Özdemir, D. (2025). Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, *37*(2), 859–965. https://doi.org/10.1007/s00521-024-10437-2 [GS Search].

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature. https://doi.org/10.1007/978-3-030-28954-6 [GS Search].

Schuch, H. S., Furtado, M., Silva, G. F. S., Kawachi, I., Chiavegatto Filho, A. D. P., & Elani, H. W. (2023). Fairness of machine learning algorithms for predicting foregone preventive dental care for adults. *JAMA Network Open*, *6*(11), e2341625–e2341625. https://doi.org/10.1001/jamanetworkopen.2023.41625 [GS Search].

Shen, S., Chen, E., Liu, Q., Huang, Z., Huang, W., Yin, Y., Su, Y., & Wang, S. (2023). Monitoring student progress for learning process-consistent knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 8213–8227. https://doi.org/10.1109/TKDE.2022.3221985 [GS Search].

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, *37*(132), 641–659. https://doi.org/10.1590/S0100-15742007000300007 [GS Search].

Silva Filho, R. L. C., Brito, K., & Adeodato, P. J. L. (2023). A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. *Expert Systems with Applications*, *221*. https://doi.org/10.1016/j.eswa.2023.119729 [GS Search].

Singelmann, L., Alvarez, E., Swartz, E., Striker, R., Pearson, M., & Ewert, D. (2020). Predicting and understanding success in an innovation-based learning course. *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*, 662–666. [GS Search].

Suaza-Medina, M., Peñabaena-Niebles, R., & Jubiz-Diaz, M. (2024). A model for predicting academic performance on standardised tests for lagging regions based on machine learning and shapley additive explanations. *Scientific Reports*, *14*(1), 25306. https://doi.org/10.1038/s41598-024-76596-3 [GS Search].

Thuy, A., & Benoit, D. F. (2024). Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, *317*(2), 330–340. https://doi.org/10.1016/j.ejor.2023.09.009 [GS Search].

Tousside, B., Dama, Y., & Frochte, J. (2022). Towards explainability in modern educational data mining: A survey. *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - KDIR*, 212–220. https://doi.org/10.5220/0011529400003335 [GS Search].

Tsiakmaki, M., & Ragos, O. (2021). A case study of interpretable counterfactual explanations for the task of predicting student academic performance. *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, 120–125. https://doi.org/10.1109/CSCC53858.2021.00029 [GS Search].

Van Petegem, C., Deconinck, L., Mourisse, D., Maertens, R., Strijbol, N., Dhoedt, B., De Wever, B., Dawyndt, P., & Mesuere, B. (2023). Pass/fail prediction in programming courses. *Journal of Educational Computing Research*, *61*(1), 68–95. https://doi.org/10.1177/07356331221085595 [GS Search].

Vieira, C. P., & Digiampietri, L. A. (2022). Machine learning post-hoc interpretability: A systematic mapping study. *Proceedings of the XVIII Brazilian Symposium on Information Systems*. https://doi.org/10.1145/3535511.3535512 [GS Search].

Vultureanu-Albişi, A., & Bădică, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 215–220. https://doi.org/10.1109/SACI51354.2021.9465558 [GS Search].

Wang, J. Z., Lan, A. S., Grimaldi, P. J., & Baraniuk, R. G. (2017). A latent factor model for instructor content preference analysis. *Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017*, 290–295. [GS Search].

Wang, Y., Zhang, Y., Liang, M., Yuan, R., Feng, J., & Wu, J. (2023). National student loans default risk prediction: A heterogeneous ensemble learning approach and the shap method. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100166 [GS Search].

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, *4*(5), e12482. https://doi.org/10.1002/eng2.12482 [GS Search].

Zanellati, A., Zingaro, S. P., & Gabbrielli, M. (2024). Balancing performance and explainability in academic dropout prediction. *IEEE Transactions on Learning Technologies*, *17*, 2140–2153. https://doi.org/10.1109/TLT.2024.3425959 [GS Search].

Zhang, S., Wang, J., Yu, S., Wang, R., Han, J., Zhao, S., Liu, T., & Lv, J. (2023). An explainable deep learning framework for characterizing and interpreting human brain states. *Medical Image Analysis*, *83*, 102665. https://doi.org/10.1016/j.media.2022.102665 [GS Search].