Journal of Applied Data Sciences
Vol. 5, No. 3, September 2024, pp. 1325-1342

ISSN 2723-6471
1325

Sustainable Educational Data Mining Studies: Identifying Key Factors and Techniques for Predicting Student Academic Performance

Murnawan[1,*], Sri Lestari[2], Rosalim Samihardjo[3], Deshinta Arrova Dewi[4]

[1,2,3]*Widyatama University, Bandung 40125, Indonesia*

[4]*INTI International University, Nilai 71800, Malaysia*

(Received: June 26, 2024; Revised: July 15, 2024; Accepted: August 30, 2024; Available online: September 4, 2024)

Abstract

This research paper presents a systematic literature review of sustainable educational data mining (EDM) studies published between 2017 and 2022 with the objective of identifying the primary factors that affect student academic performance. The purpose of this study is to provide a comprehensive analysis of sustainable EDM research and identify the most important factors that influence student performance while highlighting commonly used data mining techniques in the EDM field. The results suggest that student demographics, previous grades and class performance, social factors, and online learning activities are the most common and widely used factors for predicting student performance in educational institutions. Furthermore, Decision Trees, Naive Bayes, and Random Forests are the most frequently used categories of data mining algorithms in the studies included in the dataset. The methodology used in this study is a systematic literature review, which is a widely used technique for literature review that provides a reliable and unbiased process for reviewing data from diverse sources. The findings of this study provide valuable insights into the factors influencing student performance in educational institutions and can be used by researchers to inform future research and identify relevant factors to consider when predicting student performance.

Keywords: Data Mining Techniques, Educational Data Mining (EDM), Factors Influencing Performance, Student Academic Performance, Systematic Literature Review, Education Quality

1. Introduction

Over the past decade, the educational landscape has witnessed a growing interest in understanding the factors that influence student performance and academic success. The increasing availability of educational data has spurred the development of Educational Data Mining (EDM), a research area that leverages data mining techniques to uncover patterns and insights within educational contexts [1]. EDM aims not only to enhance educational outcomes by identifying at-risk students early but also to optimize the educational environment by providing targeted interventions and personalized learning experiences. This aligns with the broader objectives of educational institutions to foster a high-quality, inclusive educational ecosystem.

EDM has gained substantial recognition due to its capacity to process large-scale student data and extract actionable knowledge that can inform educational policies and practices. The ability to analyze such data at scale has opened up new avenues for research, with studies increasingly focusing on identifying key factors that influence student academic performance. Despite this progress, the field still grapples with a lack of consensus on the most critical determinants of student success. This ongoing debate underscores the need for a more systematic and comprehensive exploration of the factors influencing academic performance, particularly in light of the evolving educational landscape [2].

Recent advancements in EDM have introduced sophisticated data mining techniques, including machine learning algorithms, clustering, and predictive modeling, to analyze student data. These techniques have been employed to predict academic outcomes, identify patterns of student behavior, and uncover latent factors that may impact learning experiences. Despite the proliferation of research in this area, the majority of studies have focused on specific datasets or educational contexts, leading to fragmented findings that are not easily generalizable. Furthermore, while numerous studies have examined individual factors such as socioeconomic status, attendance, and engagement, there has been

***Corresponding author: Murnawan (murnawan@widyatama.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i3.347

limited integration of these factors into a holistic model that can reliably predict academic performance across diverse student populations.

A critical gap in the current body of literature is the lack of a unified framework that synthesizes the diverse factors influencing student performance and systematically evaluates their relative importance. While various studies have explored factors such as demographics, prior academic achievement, and behavioral indicators, the absence of a comprehensive review that consolidates these findings has resulted in inconsistencies in the reported determinants of student success. Additionally, there is a notable gap in the application of systematic literature reviews (SLRs) within EDM to rigorously assess the state of research and identify prevalent trends and challenges. An SLR is particularly valuable in this context, as it offers a structured approach to review, categorize, and synthesize existing studies, thereby providing a more reliable and replicable understanding of the factors that most significantly impact student performance.

This paper seeks to address the aforementioned research gaps by conducting a systematic literature review of EDM studies published between 2017 and 2022. The primary objective is to identify and synthesize the most significant factors that influence student academic performance, as well as to highlight the most commonly used data mining techniques within the EDM field. By doing so, this research aims to provide a comprehensive overview of the current state of EDM, identify key trends, and propose a more integrated framework for understanding student performance.

The contribution of this study lies in its ability to bridge the gap between fragmented research findings and establish a more coherent understanding of the factors driving student success. The systematic approach employed in this review will enhance the transparency and replicability of the findings, thereby contributing to the development of more effective educational strategies and interventions. Moreover, by identifying the state-of-the-art methodologies and highlighting areas for further research, this study will pave the way for future investigations that can build on a more solid foundation of knowledge within the EDM domain.

The systematic literature review will follow a rigorous process, including the identification of relevant studies, the extraction of data, and the synthesis of findings. This process will not only help in summarizing the existing research but also in identifying the most frequent and impactful factors associated with student academic performance. The SLR will also examine the methodological approaches used in these studies, offering insights into the strengths and limitations of current research practices in EDM.

2. Literature Review

2.1. Introduction to Educational Data Mining (EDM)

Educational Data Mining (EDM) has emerged as a pivotal research area that focuses on the application of data mining techniques to educational data. Over the past decade, this field has grown significantly, driven by the increasing availability of large-scale educational datasets and the need for data-driven decision-making in education. EDM encompasses a wide range of techniques, including classification, clustering, and regression, which are employed to analyze patterns within student data. These analyses aim to enhance the understanding of student behaviors, predict academic outcomes, and improve educational practices and policies [3]. The ability of EDM to process vast amounts of data has made it an invaluable tool for educators and researchers alike, offering insights that were previously difficult to obtain through traditional educational research methods [4].

As EDM continues to evolve, its applications have broadened to include various educational settings, from primary education to higher education. The field's versatility has enabled the development of predictive models that can identify at-risk students, optimize resource allocation, and tailor educational interventions to individual students' needs. However, despite these advancements, the field still faces challenges related to data quality, model generalizability, and ethical considerations. As the field matures, there is a growing emphasis on ensuring that EDM techniques are not only effective in specific contexts but also scalable and sustainable across different educational environments [5]. This review aims to explore the current state of EDM, identify key factors influencing student performance, and examine the techniques used to predict academic success.

2.2. Key Factors Influencing Student Academic Performance

A wide range of factors has been identified as influencing student academic performance, with demographic, academic, and behavioral factors being the most extensively studied. Demographic factors, such as age, gender, and socioeconomic status (SES), have consistently shown a significant impact on educational outcomes. For example, students from higher SES backgrounds generally have better access to resources, support systems, and educational opportunities, which contribute to higher academic achievement [6]. Gender differences have also been observed, with studies indicating that female students often outperform male students in specific subjects, such as languages and humanities, while male students may excel in areas like mathematics and science [7]. These demographic factors provide a baseline for understanding the diverse backgrounds of students and how these backgrounds can influence their academic trajectories.

In addition to demographic factors, academic and behavioral factors play a crucial role in shaping student performance. Prior academic achievement is a strong predictor of future success, with students who have performed well in previous coursework likely to continue excelling [8]. Regular attendance and active engagement in classroom activities are also critical, as they directly correlate with higher grades and better overall performance [9]. Behavioral factors, such as study habits, time management, and participation in extracurricular activities, further influence academic outcomes. Effective study habits and consistent effort outside the classroom have been linked to improved grades, while involvement in extracurricular activities can enhance social skills and foster a sense of belonging, both of which contribute to academic success [10]. Together, these factors create a comprehensive picture of the elements that contribute to student performance, highlighting the need for multifaceted approaches in educational research and intervention strategies.

2.3. Data Mining Techniques in EDM

Data mining techniques have become increasingly sophisticated in the field of EDM, with a variety of methods being used to analyze and predict student performance. Classification algorithms, such as decision trees, support vector machines (SVM), and random forests, are among the most commonly employed techniques. These algorithms categorize students based on input features into groups such as "at-risk" or "high-performing," allowing educators to intervene early and provide targeted support [11]. The use of these algorithms has proven effective in identifying students who may need additional resources or attention, thereby helping to improve educational outcomes. However, the accuracy and generalizability of these models can vary depending on the quality and quantity of the data used, as well as the specific context in which they are applied [12].

Clustering methods, such as k-means clustering and hierarchical clustering, offer another approach to analyzing student data. These methods group students with similar characteristics or behaviors, which can be particularly useful for identifying patterns and trends within student populations [13]. For instance, clustering can reveal subgroups of students who struggle with similar academic challenges, enabling more personalized instructional strategies. Additionally, regression models, including linear and logistic regression, are frequently used to quantify the relationships between various factors and student performance. These models provide valuable insights into how specific variables, such as study time or attendance, impact academic outcomes [14]. While these techniques have been widely adopted in EDM, ongoing research seeks to refine these methods and develop new approaches that can better capture the complexities of student learning and performance.

2.4. Sustainability in EDM Research

The concept of sustainability in EDM research has gained prominence in recent years, emphasizing the need for studies that are not only effective in the short term but also adaptable and scalable over time. Sustainable EDM research focuses on developing models and techniques that can be applied across various educational settings, ensuring that the insights gained are relevant and applicable in diverse contexts [15]. This approach is particularly important given the rapidly changing nature of educational environments and the increasing diversity of student populations. By prioritizing sustainability, researchers aim to create tools and methodologies that can continue to provide value as educational systems evolve, helping to ensure long-term improvements in student outcomes [16].

One of the key challenges in achieving sustainability in EDM research is ensuring the scalability of data mining techniques. Models that perform well in one educational setting may not necessarily be effective in another, due to differences in student demographics, institutional resources, and educational practices [17]. To address this issue, sustainable EDM studies emphasize the development of adaptable models that can be easily transferred and implemented in various contexts. Additionally, ethical considerations, such as data privacy and the potential for bias in predictive models, are integral to the sustainability of EDM research. Researchers must ensure that their models are transparent, fair, and designed to protect student data, which is critical for maintaining trust and ensuring the equitable application of EDM techniques [18].

Despite significant advancements in EDM, several research gaps remain that warrant further exploration. One of the most notable gaps is the lack of a unified framework that integrates the diverse factors influencing student performance into a holistic model [19]. While numerous studies have examined individual factors, such as socioeconomic status or attendance, there is a need for research that synthesizes these findings into a comprehensive understanding of how these factors interact and impact academic outcomes. Additionally, many existing studies are limited by their focus on specific datasets or educational contexts, leading to findings that may not be generalizable across different student populations or educational systems [20].

Future research in EDM should also focus on addressing the ethical implications of data mining in education, particularly in terms of ensuring fairness and transparency in predictive analytics. As data mining techniques become more sophisticated, there is an increasing need to develop models that not only accurately predict student outcomes but also do so in a way that is ethical and respects the privacy and rights of students [21]. Furthermore, the integration of emerging technologies, such as artificial intelligence and deep learning, offers exciting opportunities for enhancing the predictive capabilities of EDM models. However, these advancements must be pursued with caution, ensuring that they contribute to the development of sustainable, equitable, and effective educational practices [22].

3. Methodology

The technique of systematic literature review (SLR) is widely used for literature review and is relevant to our goal of identifying the factors that affect students' educational performance in higher education using EDM studies. The most commonly used and significant factors are identified through SLR to provide precise studies related to EDM. SLR provides an excellent system and framework for improving the quality of research articles, literature reviews, and opinions [15]. The use of clear SLR protocols aids researchers in the review process, improves the transparency of the review methodology, and enables future replication [23]. Compared to simple and unstructured literature assessment methods, SLR is considered more reliable and unbiased since it collects data from diverse sources and is larger in scope [24]. The three major phases of SLR are planning, conducting, and reporting, with each phase consisting of specific steps as outlined in figure 1. Further details about each stage are explained in subsequent sub-chapters.

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceeding. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

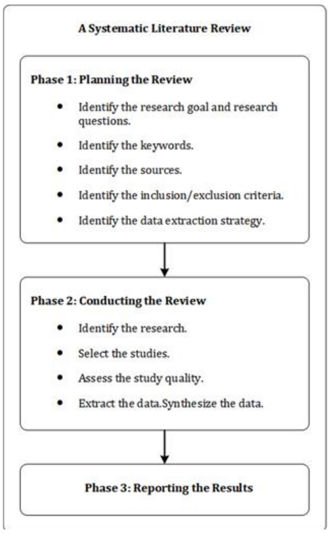

Figure 1. Overview of SLR steps [24].

3.1. Planning the Review

During the initial stage of systematic literature review, the planning phase, various essential steps must be taken into consideration. These steps include defining the research objectives and questions, determining relevant keywords, identifying relevant sources, establishing criteria for selecting and excluding studies, and devising methods for extracting and analyzing data.

3.1.1. The objective and inquiries of a research study

The objective of this article is to conduct a systematic review of the literature by utilizing the SLR approach [4], and the research questions are outlined as follows: (1) Which factors are frequently utilized to impact a student's academic performance? (2) Which data mining techniques are commonly employed to analyze and forecast a student's academic success?

3.1.2. Research keywords

The search keywords for this study were primarily based on the research questions previously mentioned. Once the search keywords were identified, a search string was created to be used with the library search engine in the following section. The search string used was as follows: ("educational data mining") AND ("influencing student performance" OR "analysis student performance" OR "predicting student performance"). Although "predicting student performance" was not specifically mentioned in the research questions, it was included in the search string as it was observed during the planning stage that many studies included this term in their titles or abstracts related to EDM and predicting student performance. The factors identified in these studies were relevant to the goals of this study at this stage.

3.1.3. Resources to be searched

We have chosen a variety of online library databases and search engines to conduct our systematic literature review, including Science Direct, Scopus, Springer, Google Scholar, IOP Science, MDPI, IEEE Xplore, and EBSCO.

3.1.4. Inclusion and exclusion criteria

Table 1 presents the inclusion and exclusion criteria that we have established for this systematic literature review. To be considered for inclusion, each study identified in the search results had to meet all of the predefined criteria.

Table 1. Inclusion and Exclusion Criteria

Inclusion	Exclusion
The study must meet the search keyword requirements.	Fails to satisfy the keyword research requirements.
It should be categorized as either educational data mining or machine learning research.	Not categorized as educational data mining research.

ISSN 2723-6471

Inclusion	Exclusion
The study must include factors that have been learned.	Does not comprise learned factors.
Full text papers must be accessible and available, and cannot be accessed via arXiv.	Full text papers are not complete or are not available, or can only be accessed through arXiv.
Review papers are excluded.	This is a paper that reviews other papers.
The study must have been published between 2017 and 2022.	Published prior to 2017.
The study must be written in English.	Written in a language other than English.

3.2. Performing the analysis or evaluation

In this section, we will elaborate on the steps we took during phase 2 of our research. These steps include identifying relevant studies, selecting appropriate ones, evaluating their quality, extracting data from them, and synthesizing the collected information.

3.2.1. Identification the research

We began the research identification phase by using the search strings specified earlier to explore online library databases. Table 3 presents the initial outcomes returned by the search engine.

Table 3. Online Library Database

Inclusion	Exclusion
Google Scholar	91
Science Direct	45
Scopus	41
Springer	36
EBSCO	17
IOP Science	16
IEEE Xplore	16
MDPI	14
Total	276

3.2.2. Study selection

The flow of information during the SLR phase is represented by PRISMA, which tracks the number of articles identified, included, excluded, and the reasons for exclusion [25]. The PRISMA flowchart, as depicted in figure 2, illustrates the process of selecting research articles that fulfill the inclusion or exclusion criteria and scrutinizing the content of the chosen articles to ensure their suitability for inclusion. Additionally, we employed automatic and semi-automated article selection techniques. The automatic selection examines the article's title, abstract, and keywords, while the semi-automatic selection reviews the full text of the article.

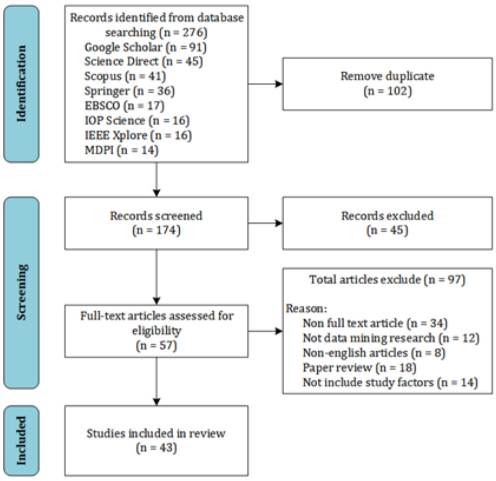

Figure 2. PRISMA flowchart.

During the study selection process, we utilized both automatic and semi-automatic methods to choose research articles that met our inclusion and exclusion criteria. Our automatic selection process resulted in the elimination of 102 duplicate results, 45 irrelevant articles that did not pertain to our data mining research topic, 12 articles that were not related to educational data mining, 34 articles that were not available in full-text or were only accessible through arXiv, 18 review papers, and 8 publications that were not written in English.

We also applied a search filter to limit results to articles published between 2017 and 2022. After this step, we were left with 57 research articles. Then, we used a semi-automatic selection method by reviewing the full text and examining the SLR material. We found that 14 of these articles did not focus on the factors that impact student performance, so they were excluded from our final dataset. In total, we included 43 research articles in our SLR. Table 4 provides a complete list of the selected studies.

Table 4. Selected Paper

Paper ID	Author(s)	Journal Name (Source)	Paper ID	Author(s)	Journal Name (Source)
SP-1	Hasheminejad and Sarvmili [23]	J. Artif. Intell. Data Min. (Google Scholar)	SP-14	Karim et al. [32]	AIUB J. Sci. Eng. (Scopus)
SP-2	Almasri et al. [7]	Arab. J. Sci. Eng. (Springer)	SP-15	Nahar et al. [33]	Educ. Inf. Technol. (Springer)
SP-3	Martínez-Abad et al. [24]	Stud. Educ. Eval. (Science Direct)	SP-16	Hussain et al. [34]	Indon. J. Electr. Eng. Comput. Sci. (Scopus)
SP-4	Triayudi and Widyarto [25]	J. Phys. Conf. Ser. (IOP Sci.)	SP-17	Injadat et al. [35]	Knowl.-Based Syst. (Science Direct)
SP-5	Yao et al. [26]	Sci. Pract. Cyber. J. (Google Scholar)	SP-18	Alturki et al. [9]	J. Inf. Technol. Educ. (Scopus)
SP-6	Saleh et al. [27]	Edukasi (Google Scholar)	SP-19	Injadat et al. [36]	Appl. Intell. (Springer)
SP-7	Bottcher et al. [13]	IEEE EDUCON (IEEE Xplore)	SP-20	Villegas-Ch et al. [37]	Sustainability (Switz.) (MDPI)
SP-8	Perchinunno et al. [28]	Soc. Indic. Res. (Springer)	SP-21	Moscoso-Zea et al. [38]	Austral. J. Eng. Educ. (Scopus)
SP-9	Asif et al. [11]	Comput. Educ. (Science Direct)	SP-22	Karthikeyan et al. [39]	Soft Comput. (Springer)
SP-10	Gowri et al. [29]	IOP Conf. Ser. (IOP Sci.)	SP-23	Queiroga et al. [40]	Appl. Sci. (Switz.) (Scopus)
SP-11	Berlilana and Mu'amar [30]	J. Digit. Mark. Digit. Curr., (Google Scholar)	SP-24	García-Jiménez et al. [18]	Sustainability (Switz.) (Springer)

SP-12	Fernandes et al. [17]	J. Bus. Res. (Science Direct)	SP-25	Wakelam et al. [41]	Br. J. Educ. Technol. (Scopus)
SP-13	Nkomo and Nat [31]	TechTrends (Springer)			

3.2.3. Assessing the quality of a research study

In order to evaluate the quality of the chosen publications, we need to address the inquiries presented in table 5 for each article in our dataset. We created a checklist based on Kitchenham's paper, which we use to appraise the quality of evidence offered by the studies we selected in the previous section.

Table 5. Assessing the Quality of A Research Study

Number	Question	Answer
QA-1	Are the study goals well-defined?	Yes/Partially/No
QA-2	Is the research sufficiently described?	Yes/Partially/No
QA-3	Does the study examine various perspectives and contexts?	Yes/Partially/No
QA-4	Are the objectives clearly connected to the conclusions?	Yes/Partially/No
QA-5	Are the findings significant?	Yes/Partially/No
QA-6	Are any negative findings reported?	Yes/Partially/No
QA-7	Do the researchers clarify the implications of any issues?	Yes/Partially/No
QA-8	Does the study enhance your knowledge or comprehension?	Yes/Partially/No
QA-9	Do the results supplement the existing literature?	Yes/Partially/No

The study utilized the answer scale from Al-Araibi et al. [5], as presented in table 6, to evaluate the quality of the selected research papers. The score of a study on this scale indicates its ability to address research questions and overall quality, with a higher score indicating better quality.

Table 6. Score of Questions in Quality Assessment Checklist

Answer	Score
Yes	1
Partially	0.5
No	0

Table 6 provides a scale of questions for each study, including a column labeled "% Max S" which displays the percentage of the maximum possible score attained by each study. The quality assessment results for the included studies are presented in Table 6.

$$\%\text{Max S} = \frac{\text{Total score for each include study}}{9} \times 100 \tag{1}$$

For example, the %Max S for SP-1 was expressed as:

$$\%\text{Max S} = \frac{7{,}5}{9} \times 100 = 83\% \tag{2}$$

The cumulative quality scores and their corresponding percentage results for all papers in our dataset are presented in table 7.

 ISSN 2723-6471

Table 7. Quality Assessment of Included Studies

Paper ID	QA-1	QA-2	QA-3	QA-4	QA-5	QA-6	QA-7	QA-8	QA-9	Scores	Percentage
SP-1	1.0	1.0	0.5	1.0	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-2	1.0	1.0	1.0	1.0	1.0	0.0	0.5	1.0	1.0	7.5	83%
SP-3	1.0	1.0	1.0	1.0	0.5	0.0	1.0	0.5	1.0	7.0	78%
SP-4	1.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5	1.0	5.0	56%
SP-5	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-6	1.0	0.5	1.0	0.5	0.5	0.0	0.5	0.5	1.0	5.5	61%
SP-7	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-8	1.0	1.0	1.0	1.0	0.5	0.0	0.5	0.5	1.0	6.5	72%
SP-9	1.0	1.0	0.5	0.5	0.5	0.5	1.0	1.0	1.0	7.0	78%
SP-10	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%
SP-11	1.0	1.0	1.0	0	0.5	0.5	1.0	1.0	0.5	6.5	72%
SP-12	1.0	0.5	0.5	0.5	1.0	0.5	1.0	1.0	0.5	6.5	72%
SP-13	1.0	1.0	1.0	0.5	1.0	0.0	0.5	0.5	1.0	6.5	72%
SP-14	1.0	1.0	0.5	0.5	0.5	0.0	0.5	1.0	0.5	5.5	61%
SP-15	1.0	1.0	0.5	0.5	0.5	0.0	1.0	1.0	1.0	6.5	72%
SP-16	1.0	1.0	0.5	0.5	0.5	0.5	1.0	1.0	1.0	7.0	78%
SP-17	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0	1.0	7.0	78%
SP-18	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-19	1.0	1.0	0.5	1.0	0.5	0.5	0.5	1.0	1.0	7.0	78%
SP-20	1.0	1.0	0.5	0.5	1.0	0.5	0.5	1.0	1.0	7.0	78%
SP-21	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-22	1.0	1.0	0.5	0.5	1.0	0.0	1.0	1.0	1.0	7.0	78%
SP-23	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%
SP-24	1.0	1.0	0.5	1.0	0.5	0.0	0.5	1.0	1.0	6.5	72%
SP-25	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-26	1.0	1.0	1.0	0.5	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-27	1.0	1.0	0.5	0.5	0.5	0.0	0.5	1.0	1.0	6.0	67%
SP-28	1.0	1.0	0.5	1.0	0.5	0.5	0.5	1.0	1.0	7.0	78%
SP-29	1.0	1.0	0.5	1.0	0.5	0.0	1.0	1.0	1.0	7.0	78%
SP-30	1.0	1.0	0.5	1.0	1.0	0.0	0.5	1.0	1.0	7.0	78%
SP-31	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%
SP-32	1.0	1.0	1.0	0.5	0.5	0.0	1.0	1.0	1.0	7.0	78%
SP-33	1.0	0.5	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.0	67%
SP-34	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%
SP-35	1.0	1.0	0.5	0.5	1.0	0.0	1.0	1.0	1.0	7.0	78%
SP-36	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%
SP-37	1.0	1.0	0.5	0.5	1.0	0.5	0.5	1.0	1.0	7.0	78%
SP-38	1.0	1.0	0.5	0.5	0.5	0.0	0.5	0.5	0.5	5.0	56%
SP-39	1.0	1.0	0.5	0.5	1.0	0.5	0.5	1.0	1.0	7.0	78%
SP-40	1.0	1.0	1.0	0.5	0.5	0.0	1.0	1.0	1.0	7.0	78%
SP-41	1.0	1.0	1.0	0.5	1.0	0.5	0.0	0.5	0.5	6.0	67%
SP-42	1.0	1.0	1.0	0.5	1.0	0.0	0.5	0.5	0.5	6.0	67%
SP-43	1.0	1.0	0.5	0.5	1.0	0.0	0.5	1.0	1.0	6.5	72%

Based on table 7, the majority of studies included in the SLR obtained scores above 4.5 out of a total score of 9, indicating a 50% threshold. Hence, all items scoring above 50% are retained in the SLR process. SP-2 received the highest score of 7.5 out of a total score of 9, corresponding to 83%. On the other hand, SP-4 and SP-38 scored the lowest, obtaining a score of 5 out of 9, which is equivalent to 56%. In this phase, all items are assessed against a quality assessment checklist and graded accordingly. The total number of papers retained at this stage is 43.

3.2.4. Data Extraction Strategy

During this step, we extracted the necessary data for our SLR study from the chosen articles based on the details presented in table 2. The primary focus of this step was to identify the factors that impact student performance, along with the data mining algorithms and techniques used by researchers in their data mining efforts. This knowledge was critical in generating valuable insights and findings that would enable us to address our research questions. Table 8 presents the outcomes of the condensed data extracted for each paper in our collection.

Table 8. Summary Data Extraction

Paper ID	Factor's Category	Data Mining Approach	Data Collection Technique
SP-1	E-learning activity of students	Classification	e-Learning system data
SP-2	The demographic information of students	Clustering	Data from the Student Information System
SP-3	The demographic information of students	Classification	Surveys about course evaluations
SP-4	The past academic performance of students in prior classes and grades; E-learning activity of students	Classification	e-Learning system data
SP-5	The past academic performance of students in prior classes and grades; The demographic information of students; Social data of students	Classification	Data from the Student Information System.
SP-6	E-learning activity of students	Association	Surveys about course evaluations
SP-7	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.; Course Evaluations	Classification	Data from the Student Information System
SP-8	Social data of students; The demographic information of students The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-9	The demographic information of students; The past academic performance of students in prior classes and grades	Classification	Data from the Student Information System
SP-10	The demographic information of students; Social data of students	Association, Clustering	Data from the Student Information System
SP-11	The demographic information of students; Social data of students	Association	Data from the Student Information System
SP-12	The demographic information of students; Social data of students	Classification	Data from the Student Information System
SP-13	E-learning activity of students	Clustering	e-Learning system data
SP-14	Social data of students; The demographic information of students; E-learning activity of students	Classification	Data from the Student Information System e-Learning system data
SP-15	Course Evaluations	Classification	Surveys about course evaluations
SP-16	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Surveys about course evaluations
SP-17	E-learning activity of students	Classification	e-Learning system data
SP-18	The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-19	E-learning activity of students	Clustering	e-Learning system data

Paper ID	**Factor's Category**	**Data Mining Approach**	**Data Collection Technique**
SP-20	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-21	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-22	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-23	E-learning activity of students	Classification	e-Learning system data; Surveys about course evaluations; Data from the Student Information System
SP-24	The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-25	The demographic information of students The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-26	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification, Clustering	Surveys about course evaluations
SP-27	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-28	The past academic performance of students in prior classes and grades.	Classification, Association	Data from the Student Information System; Surveys about course evaluations
SP-29	The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-30	Student absence behaviour	Classification	Data from the Student Information System
SP-31	Social data of students; The demographic information of students; The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-32	E-learning activity of students	Clustering	e-Learning system data
SP-33	The demographic information of students; The past academic performance of students in prior classes and grades.	Clustering	Surveys about course evaluations
SP-34	Social data of students; The demographic information of students; E-learning activity of students	Classification	e-Learning system data
SP-35	Social data of students; The demographic information of students	Classification	Data from the Student Information System
SP-36	The past academic performance of students in prior classes and grades.	Classification	Data from the Student Information System
SP-37	The demographic information of students; The past academic performance of students in prior classes and grades.	Classification, Clustering,	Data from the Student Information System
SP-38	Social data of students; The demographic information of students	Association	Data from the Student Information System
SP-39	The past academic performance of students in prior classes and grades.	Clustering	Data from the Student Information System
SP-40	Course Evaluations	Association	Surveys about course evaluations
SP-41	Course Evaluations	Association	Data from the Student Information System
SP-42	Instructor Attributes	Classification	Surveys about course evaluations
SP-43	Instructor Attributes	Classification	Surveys about course evaluations

We have identified 6 categories of factors that influence student performance by extracting data from 43 research papers. During the data collection process, we categorized each set of factors recorded on the papers. The categories can be seen in table 9.

Table 9. Factors Categories Description

Category	Description	Papers ID	Numbers of Papers
Students e-Learning activity	Data related to student engagement with the e-Learning system, including login frequency, completion of assignments and quizzes, and other related activities.	SP-1, SP-4, SP-13, SP-14, SP-17, SP-19, SP-23, SP-32, SP-34	9
The past academic performance of students in prior classes and grades.	Performance indicators from prior courses, semesters, or academic years.	SP-4, SP-5, SP-7, SP-8, SP-9, SP-16, SP-18, SP-20, SP-21, SP-22, SP-24, SP-25, SP-26, SP-27, SP-28, SP-29, SP-31, SP-33, SP-36, SP-37, SP-39	21
The demographic information of students	Demographic information about students, including age, gender, nationality, ethnicity, and other related factors.	SP-2, SP-3, SP-4, SP-5, SP-6, SP-7, SP-8, SP-9, SP-10, SP-11, SP-12, SP-14, SP-16, SP-18, SP-20, SP-21, SP-22, SP-25, SP-26, SP-27, SP-29, SP-31, SP-33, SP-34, SP-35, SP-37, SP-38	27
Instructor Attributes	Instructor-related information and evaluations of their teaching performance.	SP-42, SP-43	2
Social data of students	Social data, such as the number of friends a student has or whether they smoke.	SP-5, SP-7, SP-8, SP-10, SP-11, SP-12, SP-14, SP-16, SP-20, SP-21, SP-22, SP-26, SP-27, SP-31, SP-34, SP-35, SP-38	17
Course Evaluations	Results from course evaluation surveys that assess course clarity, satisfaction, and other relevant aspects.	SP-7, SP-15, SP-40, SP-41	4

3.3. Results the Review

In this section, we present the outcomes of our SLR analysis, where we address our research inquiries and depict the noteworthy findings that we derived from the gathered data.

3.3.1. Research distribution by categories

We categorize each research paper into one or multiple categories outlined in table 9. The categories include: (1) E-learning activity of students; (2) The past academic performance of students in prior classes and grades.; (3) The demographic information of students; (4) Instructor Attributes; (5) Social data of students; and (6) Course Evaluations.

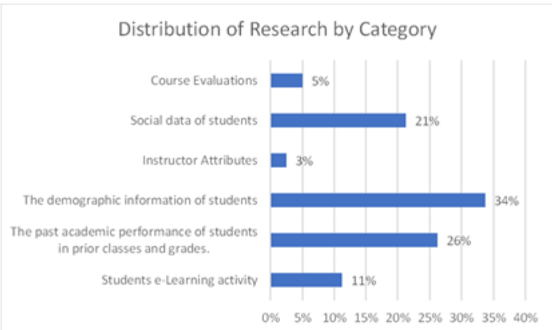

Figure 3. Distribution of research by category.

Figure 3 illustrates that the highest percentage (34%) of categorical factors predicting student performance in college is student demographics, followed by students' previous grades and class performance (26%), students' social data (21%), and student online learning activities (11%). These four categories were present in 92% of the studies.

This finding is consistent with Shahiri et al.'s [44] previous systematic literature review, which showed that CGPA and internal assessment scores were the most commonly used attributes in the EDM community to predict student performance, fitting into our top factor category.

The remaining five categories, accounting for a total of 8%, were utilized in some studies but were not frequently found in other research papers and were therefore considered as temporary factors.

3.3.2. Research Distribution by Year

According to the collected articles, the distribution of research by year reveals that the demand for educational data mining research was highest in 2018, with more than 28% of the studies being conducted in that year. There was a significant increase in interest starting from 2020 (see table 10).

Table 10. Research Distribution by Year

Year	Paper ID	Number of Papers
2017	SP-9, SP-10, SP-28, SP-40	4
2018	SP-1, SP-16, SP-26, SP-31, SP-32, SP-34, SP-35, SP-41, SP-43	9
2019	SP-5, SP-6, SP-12, SP-21, SP-42	5
2020	SP-2, SP-3, SP-7, SP-11, SP-17, SP-19, SP-20, SP-22, SP-24, SP-25, SP-29, SP-36	12
2021	SP-4, SP-8, SP-13, SP-14, SP-15, SP-18, SP-23, SP-27, SP-30	9
2022	SP-33, SP-37, SP-38, SP-39	4

3.3.3. Research Distribution by Data Collection Technique

To expand our data collection, we also considered the data collection techniques used by the researchers in the articles we obtained. We identified three main techniques: gathering data from Student Information Systems, conducting Course Evaluation Surveys, and using e-Learning system data. By summarizing the data types used by the researchers, we were able to provide an overview of the data collection techniques employed in the studies included in our dataset [42], [43].

Table 11. Summary of Data Types used by Researchers

Technique	Descriptions	Paper ID	Number of Papers
Data from the Student Information System	Information extracted from student information systems, including demographic information, admission data, and academic performance records.	SP-2, SP-5, SP-6, SP-7, SP-8, SP-9, SP-10, SP-11, SP-12, SP-14, SP-18, SP-20, SP-21, SP-22, SP-23, SP-24, SP-25, SP-27, SP-28, SP-29, SP-30, SP-31, SP-35, SP-36, SP-37, SP-38, SP-39, SP-41	27
Surveys about course evaluations	The solution is undoubtedly a survey for evaluation, usually administered after completing each course.	SP-3, SP-5, SP-15, SP-16, SP-23, SP-26, SP-28, SP-33, SP-40, SP-42, SP-43	11
e-Learning system data	Data collected from electronic learning systems.	SP-1, SP-4, SP-13, SP-14, SP-17, SP-19, SP-23, SP-32, SP-34	9

Table 11 and figure 4 display the distribution of research articles using data collection techniques. The data indicates that the most commonly used technique is the collection of Data from the Student Information System, accounting for nearly 58% of the studies. The second most utilized method was through Surveys about course evaluations (23%), followed by e-Learning system data (19%).

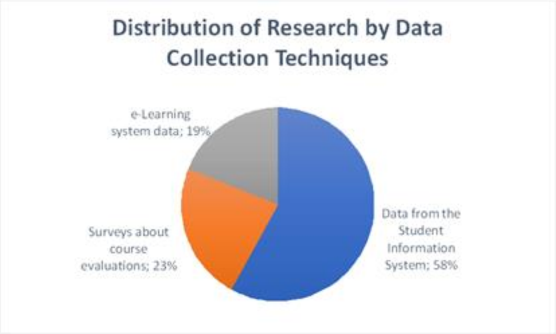

Figure 4. The distribution of research by the method of data collection.

3.3.4. Research Distribution by Data Mining Approaches

The predominant data mining techniques used in most of the research are association, classification, and clustering. A summary of the distribution of studies in our dataset using these three techniques is presented in table 11. The primary method used in data mining is classification, with almost all studies employing it to classify and predict student performance. In contrast, only four studies used association, three studies used association with classification and clustering, six studies used clustering, and three studies used clustering in combination with classification and association (see table 12). This finding aligns with previous research by Peña-Ayala, which also demonstrated that classification and clustering are the most commonly used data mining techniques in EDM research [45].

Table 12. Research Distribution by Data Mining Approaches

Data mining approaches	Paper ID	Number of Papers
Association	SP-6, SP-10, SP-11, SP-28, SP-38, SP-40, SP-41	7
Classification	SP-1, SP-3, SP-4, SP-5, SP-7, SP-8, SP-9, SP-12, SP-14, SP-15, SP-16, SP-17, SP-18, SP-20, SP-21, SP-22, SP-23, SP-24, SP-25, SP-26, SP-27, SP-28, SP-29, SP-30, SP-31, SP-34, SP-35, SP-36, SP-37, SP-42, SP-43	31
Clustering	SP-2, SP-10, SP-13, SP-19, SP-26, SP-32, SP-33, SP-37, SP-39	9

In addition, we gathered information on over 100 data mining algorithms utilized by the 43 research papers in our dataset. The top algorithms used in four or more research articles (which is over 10% of the total articles) are presented in table 13.

Table 13. Data Mining Algorithm that are Used Most Frequently

Algorithm	Frequency	Percentage
Naïve Bayes	12	28%
Decision Tree	11	26%
J48	8	19%
Random Forest	8	19%
C4.5	7	16%
Apriori	7	16%
K-Nearest Neighbours	6	14%
K-Means	6	14%
Support Vector Machine	5	12%

To simplify the analysis of data mining algorithms used in the 43 articles of our dataset, we grouped similar algorithms into seven categories. For example, ID3 and C4.5 were grouped together as decision tree algorithms. Table 14 and figure 5 present the seven algorithm categories and their frequency of use in the studies included in our dataset.

Table 14. The Most Frequently Employed Algorithms Categorized by Type

Algorithm	Frequency	Percentage
Naïve Bayes	33	77%
Decision Tree	12	28%
J48	8	19%
Random Forest	7	16%
C4.5	6	14%
Apriori	6	14%
K-Nearest Neighbours	6	14%
K-Means	5	12%
Support Vector Machine	36	84%

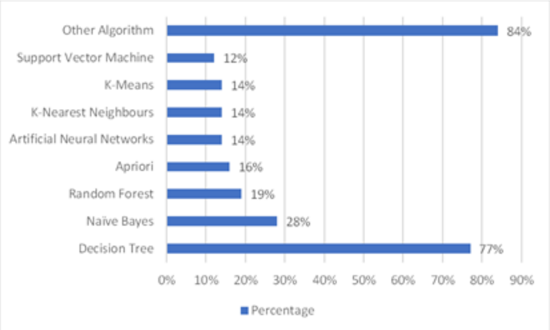

Figure 5. The algorithm used by category.

Table 14 and figure 5 illustrate that Decision Trees, Naive Bayes, and Random Forests are the most frequently used categories of data mining algorithms in the studies included in our dataset. These findings are in agreement with previous systematic literature reviews conducted by Peña-Ayala and Shahiri et al. [44], [45], which also identified decision trees and Naive Bayes as the most commonly used data mining techniques in EDM research.

4. Conclusion

This study is focused on educational data mining, with the aim of creating a standard set of factors that influence student academic performance. The main objective of this research is to identify the most common and widely researched factors that affect student performance in educational institutions, as recognized by the EDM community. Additionally, the study identifies the most commonly used data mining approaches, techniques, and algorithms to analyze and predict student performance.

The research methodology used in this study was a systematic literature review, which involved several phases and steps. The first stage of the review was to plan the review, including the development of research questions, the establishment of inclusion and exclusion criteria, and the definition of a data extraction strategy. The second stage consisted of a review step, which involved searching and identifying research articles for literature review, evaluating the quality of the selected research articles, and extracting and synthesizing data.

The findings of the study suggest that the most common and widely used factors for predicting student performance in educational institutions are student demographics, followed by the student's previous grades and class performance,

ISSN 2723-6471

social factors, and online learning activities. Moreover, the results indicate that the most commonly used data mining techniques in the field of data mining in education are Decision Trees, Naive Bayes, and Random Forests.

In conclusion, this research paper provides a systematic review of the most prevalent and most researched factors influencing student academic performance in educational institutions by the EDM community. The study also identifies the most commonly used data mining approaches, techniques, and algorithms to analyze and predict student performance. The findings of this study can be used by researchers to inform future research and to identify the most relevant factors to consider when predicting student academic performance.

For future research, scholars can utilize the key findings of this study to inform their own research, particularly regarding the most commonly researched categorical factors and data mining techniques related to student performance. These categorical factors can also be adapted to fit the context of different educational institutions, leading to the discovery of new or unique categorical factors.

5. Declarations

4.1. Author Contributions

Conceptualization: M., S.L., R.S., and D.A.D.; Methodology: M., S.L., R.S., and D.A.D.; Software: M., S.L.; Validation: M., S.L., R.S., and D.A.D.; Formal Analysis: M., S.L., R.S., and D.A.D.; Investigation: M., S.L.; Resources: S.L., R.S., D.A.D.; Data Curation: S.L., R.S., D.A.D.; Writing Original Draft Preparation: M., S.L., R.S., and D.A.D.; Writing Review and Editing: S.L., R.S., D.A.D., and M.; Visualization: M., S.L.; All authors have read and agreed to the published version of the manuscript.

4.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

4.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

4.4. Institutional Review Board Statement

Not applicable.

4.5. Informed Consent Statement

Not applicable.

4.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Education and Information Technologies,* vol. 23, no. 1, pp. 537–553, 2018.

[2] M. Wook, Z. M. Yusof, and M. Z. A. Nazri, "Educational data mining acceptance among undergraduate students," *Education and Information Technologies,* vol. 22, no. 3, pp. 1195–1216, 2017.

[3] A. Abazeed and M. Khder, "A classification and prediction model for student's performance in university level," *Journal of Computer Science,* vol. 13, no. 7, pp. 228-233, 2017.

[4] T. T. Tin, L. S. Hock, and O. M. Ikumapayi, "Educational Big Data Mining: Comparison of multiple machine learning algorithms in predictive modelling of student academic performance," *International Journal of Advanced Computer Science and Applications,* vol. 15, no. 6, pp. 633–645, 2024.

[5] A. A. M. Al-Araibi, M. N. bin Mahrin, and R. C. M. Yussof, "A systematic literature review of technological factors for e-learning readiness in higher education," *Journal of Theoretical and Applied Information Technology,* vol. 93, no. 2, pp. 500-521, 2016.

[6] D. Alboaneen, "Development of a web-based prediction system for students' academic performance," *Data*, vol. 7, no. 2, 2022.

[7] A. Almasri, R. S. Alkhawaldeh, and E. Çelebi, "Clustering-based EMT model for predicting student performance," *Arabian Journal for Science and Engineering*, vol. 45, no. 12, pp. 10067-10078, 2020.

[8] M. Alsuwaiket, C. Dawson, and F. Batmaz, "Measuring the credibility of student attendance data in higher education for data mining," *International Journal of Information and Education Technology*, vol. 8, no. 2, pp. 121–127, 2018.

[9] S. Alturki, N. Alturki, and H. Stuckenschmidt, "Using educational data mining to predict students' academic performance for applying early interventions," *Journal of Information Technology Education: Innovations in Practice*, vol. 20, no. 1, pp. 121–137.

[10] A. Ashraf, S. Anwer, and M. G. Khan, "A comparative study of predicting student's performance by use of data mining techniques," *American Scientific Research Journal for Engineering, Technology and Sciences*, vol. 44, no. 1, pp. 122–136, 2018.

[11] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers and Education*, vol. 113, no. 1, pp. 177–194, 2017.

[12] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data mining for students' disposition analysis," *Education and Information Technologies*, vol. 23, no. 2, pp. 957–984, 2018.

[13] A. Bottcher, V. Thurner, and T. Hafner, "Applying data analysis to identify early indicators for potential risk of dropout in CS students," *IEEE Global Engineering Education Conference (EDUCON)*, vol. 2020, no. April, pp. 827–836, 2020.

[14] D. Budgen, and P. Brereton, "Performing systematic literature reviews in software engineering," *Proceedings - International Conference on Software Engineering*, vol. 28, no. 1, pp. 1051-1052, 2006.

[15] M. Ezz, and A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program," *Education and Information Technologies*, vol. 25, no. 4, pp. 2733–2746, 2020.

[16] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, no. 1, pp. 19558–19571, 2022.

[17] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *Journal of Business Research*, vol. 94, no. Jan., pp. 335–343, 2019.

[18] J. García-Jiménez, J. Rodríguez-Santero, and J. J. Torres-Gordillo, "Influence of contextual variables on educational performance: A study using hierarchical segmentation trees," *Sustainability (Switzerland)*, vol. 12, no. 23, pp. 1–10, 2020.

[19] W. Gata, G. Grand, R. Fatmasari, B. Baharuddin, Y. E. Patras, R. Hidayat, S. Tohari, and N. K. Wardhani, "Prediction of teachers' lateness factors coming to school using C4.5, random tree, random forest algorithm," *Advances in Social Science, Education and Humanities Research*, vol. 258, no. 2, pp. 161–166, 2019.

[20] R. Mallett, J. Hagen-Zanker, R. Slater, and M. Duvendack, "The benefits and challenges of using systematic reviews in international development research," *Journal of Development Effectiveness*, vol. 4, no. 3, pp. 445–455, 2012.

[21] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Inf. Softw. Technol.*, vol. 55, no. 12, pp. 2049–2075, 2013.

[22] A. Liberati, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration," *PLoS Medicine*, vol. 6, no. 7, pp. 1-10, 2021.

[23] S. M. Hasheminejad, and M. Sarvmili, "S3PSO: Students' performance prediction based on particle swarm optimization," *Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 1, pp. 77–96, 2018.

[24] F. Martínez-Abad, A. Gamazo, and M. J. Rodríguez-Conde, "Educational data mining: Identification of factors associated with school effectiveness in PISA assessment," *Studies in Educational Evaluation*, vol. 66, no. 1, pp. 100844, 2020.

[25] A. Triayudi, and W. O. Widyarto, "Comparison J48 and Naïve Bayes methods in educational analysis," *Journal of Physics: Conference Series*, vol. 1933, no. 1, pp. 1-12, 2021.

[26] Y. Yao, Z. Chen, S. Byun, and Y. Liu, "Using data mining classifiers to predict academic performance of high school students," *Scientific and Practical Cyber Security*, vol. 3, no. 2, pp. 18–35, 2019.

[27] M. A. Saleh, S. Palaniappan, and N. A. A. Abdalla, "Education is an overview of data mining and the ability to predict the performance of students," *Edukasi*, vol. 15, no. 1, pp. 19–28, 2021.

[28] P. Perchinunno, M. Bilancia, and D. Vitale, "A statistical analysis of factors affecting higher education dropouts," *Social Indicators Research*, vol. 156, no. 2–3, pp. 341–362, 2021.

[29] G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational data mining application for estimating students performance in Weka environment," *IOP Conference Series: Materials Science and Engineering*, vol. 263, no. 3, pp. 1-8, 2017.

[30] Berlilana and A.Mu'amar, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 84-102, 2024

[31] L. M. Nkomo, and M. Nat, "Student engagement patterns in a blended learning environment: An educational data mining approach," *TechTrends*, vol. 65, no. 5, pp. 808–817, 2021.

[32] M. Al Karim, M. Y. Ara, M. M. Masnad, M. Rasel, and D. Nandi, "Student performance classification and prediction in fully online environment using decision tree," *AIUB Journal of Science and Engineering*, vol. 20, no. 3, pp. 70–76, 2021.

[33] K. Nahar, B. I. Shova, T. Ria, H. B. Rashid, and A. H. M. S. Islam, "Mining educational data to predict students performance," *Education and Information Technologies*, vol. 26, no. 5, pp. 6051–6067, 2021.

[34] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447–459, 2018.

[35] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, no. 1, pp. 105992, 2020b.

[36] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, 2020a.

[37] W. Villegas-Ch, X. Palacios-Pacheco, and S. Luján-Mora, "A business intelligence framework for analyzing educational data," *Sustainability (Switzerland)*, vol. 12, no. 14, pp. 1–21, 2020.

[38] O. Moscoso-Zea, P. Saa, and S. Luján-Mora, "Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining," *Australasian Journal of Engineering Education*, vol. 24, no. 1, 2019.

[39] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation," *Soft Computing*, vol. 24, no. 24, pp. 18477–18487, 2020.

[40] E. M. Queiroga, C. R. Enríquez, C. Cechinel, A. P. Casas, V. R. Paragarino, L. R. Bencke, and V. F. C. Ramos, "Using virtual learning environment data for the development of institutional educational policies," *Applied Sciences (Switzerland)*, vol. 11, no. 15, pp. 1-10, 2021.

[41] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, "The potential for student performance prediction in small cohorts with minimal available attributes," *British Journal of Educational Technology*, vol. 51, no. 2, pp. 347–370, 2020.

[42] A. R. Hananto and D. Sugianto, "Analysis of the Relationship Between Trading Volume and Bitcoin Price Movements Using Pearson and Spearman Correlation Methods", *J. Curr. Res. Blockchain.*, vol. 1, no. 1, pp. 1–19, Jun. 2024.

[43] R. A. Putawaand D. Sugianto, "Exploring User Experience and Immersion Levels in Virtual Reality: A Comprehensive Analysis of Factors and Trends," *Int. J. Res. Metav.*, vol. 1, no. 1, pp. 20-39, 2024

[44] A.M. Shahiri, W. Husain and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Computer Science*, vol. 72, pp 414–422, 2015.

[45] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4 PART 1, pp 1432–1462, 2014.