Enhancing Student Outcomes through Educational Big and Data Mining: A Systematic Literature Review

AFIF Fatima-ezzahra [1], BOUYAHIA Fatima [2], and RAFOUK Leila [1]

[1] Cadi Ayyad University, Laboratory of Didactic and University Pedagogy, Morocco
[2] Cadi Ayyad University, Laboratory of Engineering and Systems Analysis, Morocco
fatimaezzahra.afif02@gmail.com

Abstract. The world is changing at rapid rhythm, the rise of digital technologies such as big data, artificial intelligence, simulation... is creating drastic innovations and new opportunities for learning and development. This is leading to the 4th industrial revolution known as Industry 4.0, which is hugely changing the world of work. As a result, traditional teaching models are becoming obsolete, so it would be essential to rethink and customize the processes of information generation and transfer to make them more efficient and flexible. Big data has emerged as a game-changer in various industries, and education is no exception. The collection and analysis of large datasets, known as educational big data, have paved the way for innovative approaches in tracking student progress, identifying at-risk students, and enhancing teaching and learning. This data is mined and analyzed through the Educational Data Mining (EDM), a cross-disciplinary field involving computer science, education and statistics. It analyses and mines education related data to better understand students and setting which they learn. EDM focuses on developing new tools and algorithms for discovering data patterns, making data driving decision and solving various types of educational problems.

Keywords: Educational big data, educational data mining, student's performance.

1 Introduction

The significant advancements provided by technology across various fields today indicate that all aspects of life are undergoing transformation, particularly in the educational sector. The rapid growth of digital technologies in education has resulted in the generation of huge amounts of data, commonly known as educational big data [1]. This data landscape, which is vast and complex, presents both challenges and opportunities for educators and researchers who are striving to improve educational outcomes [2].

To fully utilize the potential of educational big data, EDM has emerged as a vital approach for the analysis and interpretation of these extensive datasets [3]. This paper focuses on educational big data and EDM. To perform the systematic review, the Kitchenham methodology, which includes five steps (1. Research Questions, 2. Data

Sources, 3. Keywords, 4. Inclusion and exclusion criteria, 5. Extraction), is used [4]. The following research questions (RQ) are crucial to its successful implementation:
RQ1: What are the key research areas in educational big data?
RQ2: What are the most frequently utilized methods and applications in EDM?
RQ3: What are the limitations of EDM, and what are the potential future directions?
The remainder of this paper is organized as follows: First, we describe the review methodology. Second, we introduce the relevance and foundational aspects of educational big data. Then, we provide an overview of the major components of EDM and its growth. Moreover, we will present the use of data mining in the resolution or analysis of problems that occur in educational settings. Finally, we address the limitations and challenges of adopting EDM, and provide insights for future research directions to enhance the interpretability and transparency of educational technologies.

2 Methodology

This section outlines the application of the Kitchenham methodology (Fig. 1). The process began with a search and extraction of titles, resulting a total of 70 initial articles: 25 related to educational big data and 45 associated to EDM. In the second step, potential studies were identified by analyzing the titles, leading to the exclusion of 10 educational big data studies, leaving 15 for further analysis. For EDM, 18 articles were excluded, reducing the count to 27 for the next phase. The final step involved reviewing the abstracts, introductions, and results based on specific inclusion and exclusion criteria. For educational big data, 9 studies were excluded, leaving 6 for in depth examination. Similarly, 13 EDM studies were excluded, leaving 14 for this review. The most of the analyzed studies were published in 2020 (n=6) and 2021 (n=5). The years 2023 and 2024 each have 3 studies, while 2018 and 2019 each have 1 study. The oldest year, 2017, includes 1 study.

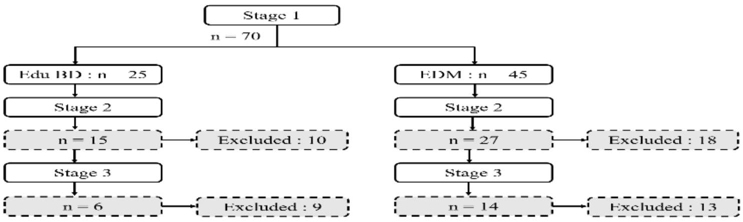


Fig. 1. Process of selection studies

3 Educational Big Data

3.1 5V's of Educational Big Data

Big data has fundamentally improved educational systems in recent years. This term refers to the extensive and diverse datasets that educational institutions collect, analyze, and use to improve teaching and learning processes, administrative functions, and student outcomes. The emergence of educational data has brought both opportunities and challenges to the field of educational big data, which is characterized by five main characteristics: Volume, Velocity, Veracity, Variety, Value [5].

Millions of schools around the world generate vast amounts of educational data, representing the number aspect. The growth rate refers to the velocity aspect. The complexity of educational big data comes from the inclusion of many entities (students, teachers, administrators) and connections (teacher-student, classmate, friend), making educational big data systems difficult due to issues like name confusion (teachers and students) and data redundancy (duplication). To overcome these challenges, however, it is vital to ensure the veracity of educational big data. The different characteristics of educational big data include behavioral aspects (fatigue, concentration, joy, surprise), life events (shopping, activities), personal information (genre, age, ethnicity, birth date, language, province, family), and learning behaviors (fraction completed, fraction spent, fraction played and fraction pause) [6]. These are just a few examples. The variety of features of educational platforms such as intelligent tutorial systems and MOOC platforms, tutorial transcripts, and reading materials include students' click-stream data [7]. The value of educational big data is one of its key features which understands the behavior or pattern generated by the data. This paves way for a predictive model generation that provide insights to personalize learning, identifying struggling students, refine teaching methods and improve overall educational outcomes. Educational institutions can make data-driven decision that enhance the effectiveness and efficiency of the learning process by understanding and utilizing these patterns.

3.2 Educational Big Data Applications

The use of big data in education improves personalized learning, teaching, and assessment as compared to traditional methods. This section explores the most beneficial big data applications [5] [7] [8]:

A. Enhanced learning experience: Educational big data seeks to improve the learner’s experience and educator efficiency by aligning the teaching and learning environment with learner’s capabilities and resources. By aggregating educational data from a large number of students and analyzing it using data mining techniques, models can be built to design adaptive learning environments that tailor the educational experience and suggest additional support, thereby boosting success and accessibility in education.

B. Improved quality of education: Educational big data is used to make informed decisions, with decision-makers relying on relevant data obtained from data analysis. They are concerned about the effective outcomes of schools to assess learners' learning success rates, patterns, difficulties, and academic achievement. Throughout the learning process, time-stamped inputs and behaviors of learners are stored. This helps in tracking learners' progress and future learning outcomes, as well as drop-outs. It establishes a new model for stakeholders to select best practices that simply learning processes and ensure that organizations are accountable for learners’ attritions, satisfaction, and failures.

C. Enhanced market analysis: By using educational big data, organizations can enhance their market analysis by tracking and analyzing their success rates, achievements, and weaknesses compared to benchmarked institutions. This data also enables the evaluation of business performance, key indicators, and academic accreditation. It helps present evidence of success, identify growth opportunities, and adopt a more strategic and innovative approach. Additionally, organizations can compare their learners, teachers, and curriculum results with those of similar institutions, gaining valuable insights for future enhancements and strategic planning.
E. Predictive teaching and assessment strategy: Using educational big data allows teachers to receive immediate, objective feedback on their course content and the effectiveness of their teaching and assessment methods. By analyzing learners' abilities and knowledge levels, teachers can monitor the learning process to detect weaknesses and potential failure risks early. Educational big data also helps identify and monitor defects in programs, courses, and content, providing data analysis and in-depth insights to enhance the curriculum. It supports personalized instruction, prompt formative assessments, active engagement in learning, and collaborative learning. In addition, it predicts learners' success by using their personal histories as indicators of future performance.

4 Educational Data Mining

4.1 Introduction

Data Mining is a powerful tool that can discover valuable information by analyzing data from various angles or dimensions, organizing that information, and summarizing the relationships found within the database [9]. This process ultimately supports better decision-making. In large datasets, data mining identifies useful patterns and trends, including descriptions, classification, clustering, forecasting, association, among others.

Data mining techniques are crucial in the education system. These techniques enhance learning and efficiency by predicting student outcomes, a task that was previously impossible using traditional methods. The rise of data mining in education has led to the development of EDM, which extracts valuable insights from large educational databases to forecast students' future performance. Through various activities, EDM helps students improve their academic outcomes.

4.2 Goals of Educational Data Mining

Below are the main targets set for using EDM [10]:
A. Predict Student Behavior: This goal can be achieved by developing student models that characterize and categorize a student's features or states, integrating characteristics such as awareness, attitudes, learning motivation, knowledge, and meta-cognition.
B. Enhance Knowledge Models: Utilize EDM techniques to identify, analyze, and refine models that represent educational content. This involves examining the complex

relationships between different knowledge areas and uncovering patterns that improve the organization, delivery, and understanding of educational material. By doing so, we can develop more precise and effective models that better reflect the complexities of learning and knowledge acquisition.

C. Evaluate Learning Support Systems: Analyze the impact of various educational tools to determine and optimize the most effective methods for improving student learning and academic outcomes.

D. Advance Educational Research: Develop and integrate advanced student models and innovative educational methods to push the boundaries of scientific understanding education.These models provide deeper insights into how students learn, identify key factors that influence their academic success, and uncover new patterns in learning behavior, thereby supporting the development of more effective educational theories and practices.

4.3 Educational Data Mining Settings

Over recent years, EDM has gained significant attention from researchers across various global disciplines, involving different groups of users or participants.

The data is produced by various distributed sources, including both structured and unstructured data. This data primarily originates from two sources [10] [11]: offline and online. Offline data comes from traditional and modern classrooms, interactive teaching and learning environments, learner information, student attendance, emotional data, course details, and academic records from institutions. Online data, on the other hand, is derived from geographically dispersed educational participation, distance education, web-based learning, and computer-assisted collaborative learning through social networking sites and online forums. The application of data mining in an educational system is illustrated in Figure 2 [12]. It is generally referred to as knowledge discovery in databases. In this process, knowledge is extracted or mined from vast amounts of educational data, including data from teachers, resources, students, alumni, etc. Academics and educators use EDM to design, plan, build, and maintain educational systems to improve student performance. Students interact directly with these systems, generating data that serves as input for EDM. The system provides students with recommendations and uncovers new insights for educators by employing a range of data extraction methods, including clustering, classification, and pattern comparisons.

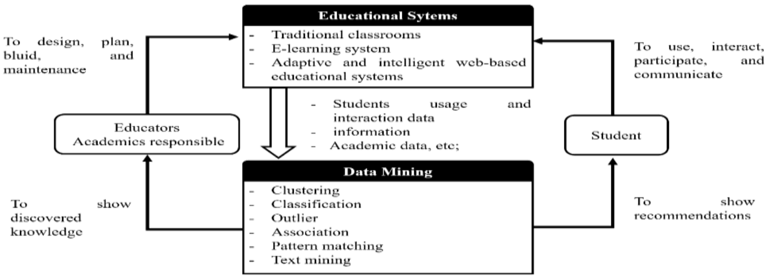


Fig. 2. The cycle of data mining in the educational system

4.4 Educational Data Mining applications

EDM has a variety of applications in the education industry, including [13]:

A. Data analysis and visualization: Data analysis and visualization are integral components of educational data mining, designed to extract meaningful insights from complex datasets and support informed decision-making in educational settings. Through statistical analysis, educators can monitor and evaluate various aspects of student behavior, such as course engagement, interaction patterns, resource utilization, and learning sequences. This analysis generates detailed summaries and reports that highlight key metrics, including session durations, content access frequency, and participation in discussions, which are crucial for understanding and improving student learning experiences. Visualization further enhances this process by translating statistical data into intuitive graphical representations, allowing educators to easily detect trends, identify outliers, and understand the overall learning trajectory of students. By combining these techniques, educators can gain a holistic view of student progress, identify areas for intervention, and tailor instructional strategies to better meet the needs of learners, ultimately leading to more effective and personalized education [14].

B. Predicting student performance: Predicting student performance is a central application of EDM, leveraging various techniques, such as regression, classification, clustering, to estimate future outcomes like grades, knowledge, and overall academic success [14]. Regression analysis, which models the relationship between dependent and independent variables, is commonly used to forecast numerical values like test scores and academic achievements. On the other hand, classification techniques categorize students based on predictive features, such as previous academic performance and demographic factors. A wide array of data mining methods, including neural networks, Bayesian networks, decision trees, and rule-based systems, have been applied to predict student outcomes. These methods analyze data from various sources, such as learning management systems, to generate models that can accurately forecast student success, identify at-risk students, and provide insights for personalized education strategies. Comparisons of different algorithms, such as Nearest Neighbor algorithm (IB1), decision trees, and logistic regression, have shown varying levels of accuracy in predicting performance, with some methods excelling in specific contexts, such as predicting dropout rates or final grades [15] [16]. Ultimately, the use of EDM in predicting student performance enables educators to make data-driven decisions, enhancing the effectiveness of teaching and learning processes.

C. Detecting Undesirable Behavior Among Students: Detecting undesirable student behaviors in EDM involves identifying students who exhibit problematic actions such as lack of motivation, cheating, misuse, dropping out, or academic failure. Techniques like classification and clustering are employed to predict and prevent these behaviors early on, allowing timely interventions. Decision trees, neural networks, support vector machines, and Bayesian networks are commonly used to detect behaviors such as low performance, dropout risks, and irregular learning patterns. Clustering methods, such as Kohonen nets and outlier detection, help uncover atypical behaviors, while association rule mining and latent response models are used to diagnose learning problems and detect system misuse. These methods enable educators to address student issues proactively, enhancing the over-all learning experience and preventing negative outcomes [17].

D. Grouping Learners: Grouping and categorizing learners in EDM involves using clustering and classification techniques to identify patterns among students with similar characteristics, such as learning styles, personality traits, and academic performance. Clustering algorithms, such as K-means, hierarchical clustering, and model-based clustering, group students based on their shared attributes, enabling the creation of personalized learning frameworks. These frameworks support adaptive content delivery, community learning, and targeted interventions. Studies have shown that clustering before classification can improve prediction accuracy, such as predicting students' final Grade Point Average (GPA), and can reduce errors like root mean squared error (RMSE) [16]. These techniques help educators and developers tailor educational experiences to meet the diverse needs of students, ultimately enhancing learning outcomes and supporting effective group learning.

E. Student modeling: Student modeling in EDM involves creating cognitive models that represent a student's knowledge, skills, learning behavior, and other characteristics such as motivation, satisfaction, and learning styles. By employing various data mining techniques like Bayesian networks, clustering, classification, and sequential pattern mining, these models can predict a student's knowledge level, learning progression, and even behavioral patterns in intelligent tutoring systems. These models also help to infer unobservable learning variables, identify learning styles, and generate personalized feedback or advice [18]. Additionally, advanced techniques like logistic regression, Markov decision processes, and fuzzy systems contribute to accurately predicting student responses, constructing adaptive learning environments, and improving the overall learning experience by tailoring educational resources to individual needs.

F. Planning & scheduling: aim to optimize the educational process by efficiently organizing future courses, resource allocation, and curriculum development. Techniques like association rule mining, decision trees, and clustering are employed to enhance course planning, predict enrollment trends, and support decision-making in higher education institutions. These methods help in analyzing student course preferences, predicting the success of curriculum revisions, and even automating course scheduling to meet the needs of both students and educators. By leveraging these data-driven insights, educational institutions can improve the quality and efficiency of their academic offerings and administrative processes [14].

4.5 Educational Data Mining methods

EDM employs a specific set of data mining methods according to the application and objective for which they are used in the EDM process (Table 1.). These methods can be classified as shown (Fig. 3) [19].

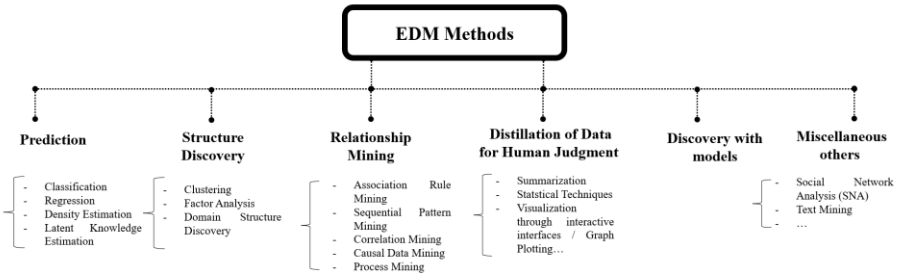
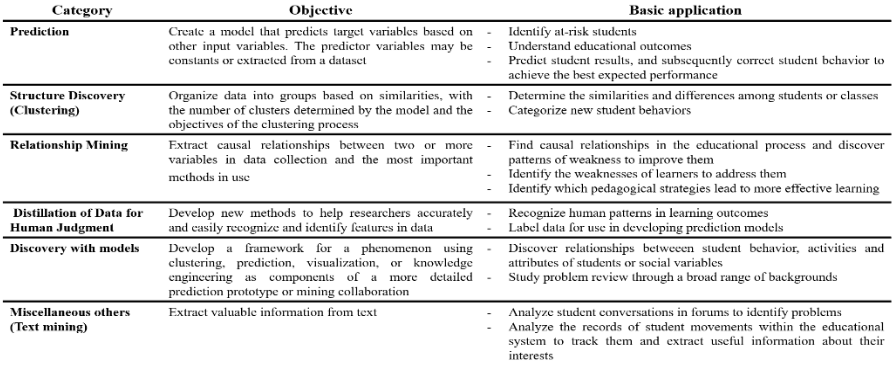


Fig. 2. Classification of EDM methods

Table 1. Objectives and applications of EDM methods



Category	Objective	Basic application
Prediction	Create a model that predicts target variables based on other input variables. The predictor variables may be constants or extracted from a dataset	- Identify at-risk students - Understand educational outcomes - Predict student results, and subsequently correct student behavior to achieve the best expected performance
Structure Discovery (Clustering)	Organize data into groups based on similarities, with the number of clusters determined by the model and the objectives of the clustering process	- Determine the similarities and differences among students or classes - Categorize new student behaviors
Relationship Mining	Extract causal relationships between two or more variables in data collection and the most important methods in use	- Find causal relationships in the educational process and discover patterns of weakness to improve them - Identify the weaknesses of learners to address them - Identify which pedagogical strategies lead to more effective learning
Distillation of Data for Human Judgment	Develop new methods to help researchers accurately and easily recognize and identify features in data	- Recognize human patterns in learning outcomes - Label data for use in developing prediction models
Discovery with models	Develop a framework for a phenomenon using clustering, prediction, visualization, or knowledge engineering as components of a more detailed prediction prototype or mining collaboration	- Discover relationships betweeen student behavior, activities and attributes of students or social variables - Study problem review through a broad range of backgrounds
Miscellaneous others (Text mining)	Extract valuable information from text	- Analyze student conversations in forums to identify problems - Analyze the records of student movements within the educational system to track them and extract useful information about their interests

5 Educational Big Data & Educational Data Mining Challenges

Despite the potential advantages of big data and EDM, there are also challenges and limitations to consider [5] [12] [20] [21]. These challenges include:

A. Data quality control: Educational data can be incomplete, inaccurate, or lacking crucial information, which complicates the process of extracting meaningful insights.

B. Privacy and security: EDM and educational big data involve managing sensitive information about students and educators, which raises important privacy and security issues. It is crucial to ensure that the collection, storage, and use of educational data are conducted in a manner that safeguards privacy and security.

C. Storage: Datasets can require significant assets to store, making it a challenging task to manage volumes of educational data.

D. Formatting and Data Cleaning: are basic steps in the data management process that require advanced technical skills due to the complexity and scale of big data. These processes ensure that data is accurate, consistent, and ready for meaningful analysis, which is essential for deriving valuable insights and making informed decisions.

E. Platform Integration: poses significant challenges, including managing diverse data formats and structures across different systems, which can lead to data redundancy and inconsistencies. These challenges often result in redundant efforts, increased costs,

and potential resource wastage as departments work to align and unify their data sources effectively.
F. Technical expertise: One challenge in EDM is the necessity for advanced technical skills. It involves sophisticated data analysis and machine learning methods, which demand a high degree of technical expertise and knowledge.
G. Interpreting results: Another difficulty in EDM is the interpretation of analytical results. The outcomes can be intricate and challenging to decipher, making it hard to derive clear and actionable insights.

6 Conclusion

The integration of EDM and educational big data into the teaching and learning process holds significant potential for revolutionizing education. As traditional teaching models become increasingly inadequate, the application of EDM offers a pathway to more efficient and flexible educational practices. By analyzing large amounts of educational data, EDM can enhance our understanding of student behavior and learning environments, ultimately leading to data-driven decisions that improve educational outcomes. This review highlights the transformative impact of EDM on education, showcasing its applications in predicting academic performance, recommending personalized learning pathways, detecting undesirable behaviors, and modeling student profiles. The insights gained from this research underscore the necessity of leveraging advanced data analytics to meet the evolving demands of modern education.

To further optimize educational practices, we propose developing a teaching model that integrates EDM techniques with Lean Management tools. By adopting this dual approach, we can enhance educational efficiency and better align teaching strategies with student and industry needs. This innovative model aims to create a dynamic and responsive educational environment that fosters continuous improvement and maximizes learning outcomes.

7 References

1. Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020b). Big data in education : a state of the art, limitations, and future research directions. International Journal Of Educational Technology In Higher Education, 17(1).
2. Lesjak, D., & Kohun, F. (2021b). BIG DATA ANALYTICS IN HIGHER EDUCATION. Issues In Information Systems.
3. Barbeiro, L., Gomes, A., Correia, F. B., & Bernardino, J. (2024b). A Review of Educational Data Mining Trends. Procedia Computer Science, 237, 88-95.
4. Kitchenham, B., & Brereton, P. (2013b). A systematic review of systematic review process research in software engineering. Information And Software Technology, 55(12), 2049-2075.
5. Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational Big Data : Predictions, Applications and Challenges. Big Data Research, 26, 100270.

6. Zhang, X., Sun, G., Pan, Y., Sun, H., He, Y., & Tan, J. (2018). Students performance modeling based on behavior pattern. Journal Of Ambient Intelligence And Humanized Computing, 9(5), 1659-1670.
7. Raza, N. M., RKayani, N. H. U., Malik, N. A., Gul, N. M., & Suleman, N. A. (2023). TRENDS AND APPLICATIONS OF BIG DATA IN EDUCATION. Pakistan Journal of Science, 75(02), 345 352.
8. Hazra, S., & Ganguli, S. (2021b). Big Data and Its Application in Smart Education during the COVID-19 Pandemic Situation. Dans CRC Press eBooks (p. 187-201).
9. Fuseini, I., & Missah, Y. M. (2024). A critical review of data mining in education on the levels and aspects of education. Deleted Journal, 1(2), 41 59
10. Bachhal, P., Ahuja, S., & Gargrish, S. (2021). Educational Data Mining : a review. Journal Of Physics Conference Series, 1950(1), 012022.
11. M. H. Qasem, R. Qaddoura and B. Hammo, "Educational Data Mining (EDM): A Review," Conference New Trends in Information Technology- (NTIT), 2017.
12. Mehra, C., & Agrawal, R. (2020). Educational data mining approaches, challenges and goals : A review. JIMS8I - International Journal Of Information Communication And Computing Technology, 8(2), 442-447.
13. Alshareef, F., Alhakami, H., Alsubait, T., & Baz, A. (2020). Educational Data Mining Applications and Techniques. International Journal Of Advanced Computer Science And Applications, 11(4).
14. Sharma, P., & Sharma, S. (2020). DATA MINING TECHNIQUES FOR EDUCATIONAL DATA : a REVIEW. International Journal Of Engineering Technologies And Management Research, 5(2), 166-177
15. Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, M. A. M. (2024b). Analyzing students' academic performance using educational data mining. Computers And Education Artificial Intelligence, 7, 100263.
16. Ampadu, Y. B. (2023). Handling Big Data in Education : A Review of Educational Data Mining Techniques for Specific Educational Problems. AI Computer Science And Robotics Technology, 2.
17. Mourabit, I. E., Jai-Andaloussi, S., & Abghour, N. (2021). Educational Data Mining Techniques for Detecting Undesirable Students' Behaviors and Predicting Students' Performance : A Comparative Study. Dans Advances in intelligent systems and computing (p. 163-170).
18. M. V. Amazona and A. A. Hernandez, "Modelling Student Performance Using Data Mining Techniques," in Proceedings of the 2019 5th International Conference on Computing and Data Engineering, May 2019, pp. 36–40,
19. Aleem A, Gore MM. Educational data mining methods: A survey. In: 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT). Gwalior, India: IEEE; 2020. pp. 182-188
20. Debang, M., & Hassan, B. U. (2023b). Educational Data Mining : Prospects and Applications. Authorea (Authorea).
21. Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining Big Data in Education : Affordances and Challenges. Review Of Research In Education, 44(1), 130-160.