Exploring Learning Patterns: A Review of Clustering in Data-Driven Pedagogy

Rohma Qadir[1], Areej Fatemah Meghji[1*], Urooj Oad[1], Veena Kumari[1]
[1]Department of Software Engineering, Mehran University of Engineering and Technology Jamshoro, Pakistan
* **Correspondence**: Areej Fatemah Meghji, areej.fatemah@faculty.muet.edu.pk

Citation | Qadir. R, Meghji. A. F, Oad. U, Kumari. V, "Exploring Learning Patterns: A Review of Clustering in Data-Driven Pedagogy", IJIST, Vol. 5 Issue. 4 pp 831-846, Dec 2023
Received | Dec 02, 2023 **Revised |** Dec 04, 2023 **Accepted |** Dec 18, 2023 **Published |** Dec 31, 2023.

Educational institutes amass and retain extensive amounts of data including records of student attendance, test scores, exam results, and performance statistics. Extracting insights from this data can provide valuable information to educators and policymakers. The rapid expansion of educational data underscores the need for sophisticated algorithms to process such vast quantities of information. This challenge led to the emergence of the field of educational data mining (EDM). Clustering is a popular approach within EDM that can find hidden patterns in data. Numerous studies in EDM have concentrated on applying diverse clustering algorithms to educational attributes. This paper presents a comprehensive literature review focusing on 43 papers spanning between 2013 to 2023 on the use of clustering algorithms and their effectiveness within the realm of EDM. The review indicates that K-means clustering has been utilized extensively in the reviewed literature with 29 of the 43 reviewed papers using K-means clustering in their analysis. It was also uncovered that cluster-based analysis majorly focuses on analyzing student performance in a course or in a degree program closely followed by clustering students based on class of learners. Insights are deduced from the reviewed literature highlighting the focus of current research and potential directions for the future.

Keywords: Clustering, Educational Data Mining (EDM), K-Means, Student Performance, Class of Learners.

Author's Contribution.
Rohma Qadir: Writing – original draft, methodology, visualization, literature review, result reporting. Areej Fatemah Meghji: methodology, writing – editing and review, validation, result reporting. Urooj Oad: Literature review, data extraction. Veena Kumari: Literature review, data extraction.

Conflict of interest.
The authors declare they have no conflict of interest for publishing this manuscript in IJIST.

Project details.
This research was not part of any project.
Acknowledgment. NIL

Introduction:

Educational Data Mining (EDM) is a sub-field of data mining that focuses on extracting useful information and patterns from large educational records [1]. EDM analyzes raw data emerging from educational institutes and converts it to knowledge that has the potential to impact educational outcomes [2]. The use of EDM has not only proved efficient at handling massive amounts of educational data but it has also helped educators better analyze and thus understand student performance. Several methods are employed to undertake EDM.

Clustering is a popular method for analyzing student behavior and learning patterns. It involves grouping students based on similar patterns, characteristics, or behavior to analyze some aspect of their behavior or performance [3]. This analysis can focus on various parameters including internal assessment factors such as CGPA, test score, mid-exam score, final examination marks, or external factors such as participation in extra circular activities [2][3]. Among others, EDM scrutinizes student descriptive, learning, attitudinal, and behavioral data to analyze and cluster similar students to discover the patterns that set one group of students apart from the other group [4]. The goal of cluster-based analysis is to group students based on emerging patterns and then use the derived knowledge to develop strategies to guide students in obtaining academic excellence. Educators have used cluster-based analysis to provide interventions to students exhibiting subpar performance, help devise curriculums, and strategize employment opportunities [5][6].

This analysis helps educators better understand how different sets of students behave and learn [7]. It also allows educators to design pedagogical policies specifically targeting each category of students. Cluster-based analysis can be carried out at the subject, semester, or degree level [4]. Research focusing on clustering different facets of student data has gained momentum in recent years. Many educational institutes had to switch to and rely on digital platforms during the COVID-19 pandemic. This led to an additional influx in the educational data being generated, creating opportunities to experiment and analyze educational content. This paper aims to explore the current state-of-the-art to find the educational areas being explored using clustering methods.

The rest of this paper is organized as follows: the review search procedure has been presented in section II highlighting the targeted knowledge sources, and the inclusion and exclusion criteria. A summary of the findings of the review has been presented in section III, followed by a discussion and answer to the review questions. The last section presents the conclusion of the paper.

Objectives:

The main aim of this research is to review the scope of clustering within the context of EDM. An attempt has been made to present a comprehensive review of the popular clustering algorithms being utilized in EDM with an emphasis on discovering the educational problems these algorithms aim to resolve. The objectives of this review are as follows:

- To identify the educational issues being targeted in cluster analysis.
- To highlight the factors/variables being used.

Material and Method:

As depicted in Figure 1, this systematic literature review follows the Kitchenham criteria to conduct the review [8]. The literature review framework, also popularized as the Kitchenham systematic literature review process, is essentially broken down into three major steps. The first phase of the review is referred to as the Plan phase. This phase consists of devising the research questions, and the formulation of the inclusion and exclusion criteria. Formulating the research questions is a critical step of the review as i) these questions serve as the roadmap for the review process, and ii) the entire purpose of the review is to provide answers to the research questions posed in this step. The identification of the knowledge sources for the review is also undertaken

during this step. The current review focuses on papers acquired through five knowledge sources: IEEE Xplore, ACM Digital Library, Journal of Educational Data Mining, Science Direct, and Google Scholar.

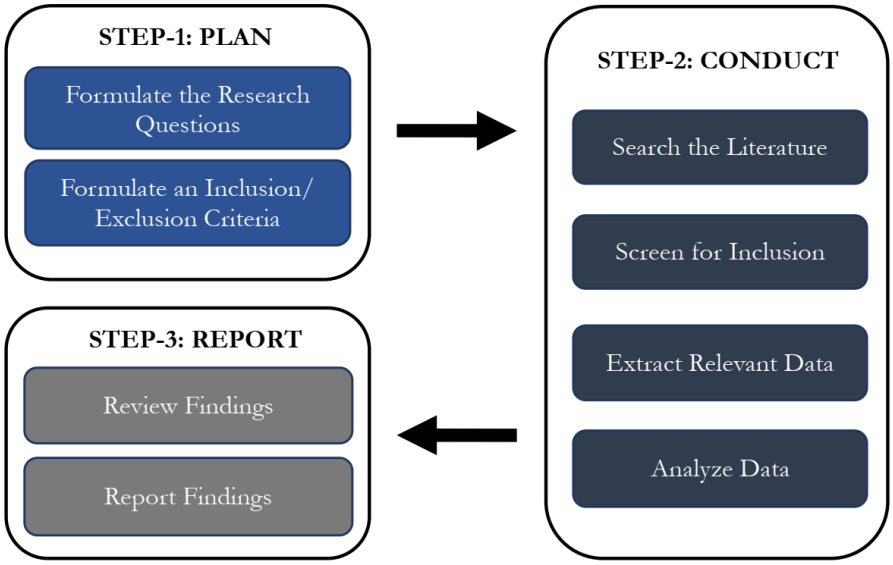

Figure 1. Systematic Literature Review Process

The Conduct stage of the review focuses on the actual review. This includes reading the papers, screening them for inclusion/exclusion, and extracting the desired information needed to answer the review questions formulated in Step 1. The last phase of the review is the Report phase. This step presents the findings of the review in terms of the answer to the devised review questions.

The Kitchenham review process is a widely used structured approach to conduct reviews in domains of software engineering. The established inclusion/exclusion criteria aid in a thorough coverage of the literature. As all the drawn conclusions are based on the evidence derived during the review, this approach towards the review offers a robust foundation for drawing inferences and making decisions.

Research Questions:

To conduct any literature review, an important step is the formulation of the research questions. The research questions proposed for this review are:

RQ1: What educational problems are being analyzed using cluster analysis?
RQ2: What are the techniques frequently selected for cluster analysis in EDM?
RQ3: Which parameters/attributes are considered for clustering-based educational research?

Search Strategy:

The following search strings were constructed for the extraction of papers: “education”, “student performance”, “learning”, “clustering” “clustering algorithm”, “clustering techniques”, “clustering approach”, and “educational data mining”. The Boolean operators AND and OR were used to join the search strings.

Inclusion Criteria:

The criteria for inclusion were established to include only papers that were relevant to the current review. The inclusion criteria set up for this literature review were:

- Only conference papers and journal articles would be included.
- Worldwide coverage of research focusing on the selected search criteria.
- Education data mining analysis focusing on clustering.
- The research paper addresses the research questions.
- The research is conducted between the years 2013 – 2023.

Exclusion Criteria:

The exclusion criteria were constructed to eradicate papers not relevant to the review process. The exclusion criteria set up for this literature review comprised of:

- Papers written in a language other than English.
- The paper's context other than the education population will not be considered.
- Papers not in the timespan of 2013 – 2023.

Range of Review Papers:

The review focused on studies from 2013 to 2023. The statistics for the knowledge sources for the studies and the resulting papers from each source have been presented in Table 1.

Table 1. Selection of Papers

No	Data Source	Primary Selection	Final Selection
1	IEEE Xplore	45	23
2	ACM Digital Library	17	4
3	Journal of Educational Data Mining	3	1
4	Science Direct	30	8
5	Google Scholar	24	7

Results and Discussion:

A summary of the findings of the review process has been presented in Table 2. An attempt has been made to highlight the educational problem or specific domain targeted for cluster-based analysis. The clustering algorithm used to tackle the issue has also been presented along with the data attributes used in the analysis. The findings of each paper have also been summarized.

Table 2. Review of the Literature

Ref.	Problem/Objective	Algorithm/Tool	Dataset/Data Source	Findings
[1]	Enhance educational process mining by making models clear and easy to understand	Expectation Maximization (EM) - WEKA	Data of 84 undergraduate students studying in a university in North Spain for a Psychology degree using Moodle 2.0.	The Heuristic net of failing students. If new students follow the same behavioral pattern, they will risk failing the course
[2]	Comparative analysis of K-means methods through a case study considering various parameters	K-means - MATLAB	Data of 455 graduate students between the years 2005 and 2009. Clustering through attributes of GPA, Program, and Study Periods.	Two silhouette comparisons with k=4 and k=5, results show that k=4 is a satisfactory value for clusters
[3]	Analyzing the performance of students while submitting assignments, using different features to observe variance in clusters	K-means	Data of Functional Programming course for undergraduate students of McGill University Canada. Features of grade, time spent, number of errors, and assignment submission collected	Students who work in the morning achieve higher grades

			through Zoom and in-person meetings	
[4]	Establish a prediction algorithm by using both classification and clustering methods combined	K-means, Support Vector Machine, Naïve Bayes, and Decision Tree	Student demographics, academic, behavioral features, and extra-curricular features in different disciplines considering higher educational institutes of Kerala, India.	Clustering algorithms perform better in conjunction with the decision tree and neural network approaches.
[5]	Provide a review of clustering from 1983 to 2016	Various clustering algorithms	Various features	Clustering techniques provide pertinent insights into variables relevant to the separation of clusters.
[6]	Analyze student performance to help decrease drop-out ratios	K-means	Data of students studying bachelor's program in the department of Computer Application B. J. College, Pune using mid, final, and assignment score	The majority of students have medium performance in the course.
[7]	Use clustering to visualize and refine a learning environment	Hierarchal method (Ward method of R cluster)	Math learning environment with 69 participating students. 454 sessions ignoring online sessions less than 5 minutes, 1030 total exercises, and students' performance in predicting while attending those full exercises.	Visualization of student learning activities. Individual students have different learning patterns that do not overlap.
[9]	Understand the COVID-19 impact on e-learning systems and also understand EDM importance in e-learning systems based on student performance	K-means, linear regression, Naïve Bayes, random forest, MLP	Not applicable.	The global e-learning market size has increased around 32.1 billion dollars and is expected to grow to 325 billion dollars by 2025
[10]	Monitor learning styles in a hybrid learning environment	K-means, X-means, K-medoids	Record of 44 students comprising of activity scores for pre-test, post-test, and time for pre-test, and post-test	The learning success of online learners is lower than onsite learners

			from the University of Phayao	
[11]	Understand the performance of the university students with data mining methods	X-Means Clustering	210 university students in the batch of years 2007-2008 and 2008-2009 considering program information technology B.E of NED University, Karachi, Pakistan.	Proved the claim that a pragmatic policy can be designed to find students who struggle in their studies
[12]	Map students using K-mean to present the hidden pattern and classify students-based GPA of certain subjects	K-means	Data of 306 students from the program of Industrial Engineering in Indonesian Islamic University.	Clever and active student groups take the percentage of 47.75% in clusters. Below average 33.33% and the least student group with 20.91%.
[13]	Analysis of course scores of college students	K-means	Data of 68 students of the Mechatronics program for the "Intelligent Control System" course.	Teachers should pay attention to Knowledge Point 1 and deeply explain Knowledge Points 2 and 4.
[14]	Understand university-level student skills based on academic course results	Decision Tree, Regression analysis, and K-K-means	The dataset comprised of 109 university students' result data	Clustering with K-means performs well and can be combined with other EDM approaches for higher performance.
[15]	Understand the quality analysis of the English learning method	K-means	College English Teaching - (CET-4) scores in which 588 data objects were considered from Sichuan, China	239 instances result in 40% proportion of the total indicating good and excellent grades in English; 80% of student's admission scores are in the range of 400 to 500.
[16]	Determine the possibility of using Moodle Courses and the performance of students	K-means	Course data from 2015 to 2019 with 12 weeks of study of Electronics, Telecommunications, and Information	The percentage of students (at-risk cluster) who did not complete

			Technologies at the Politehnica University Timisoara.	activity in the 14th week was 86%.
[17]	Develop a tool for examining clusters and graduate students' profiles.	K-means	2000 records (Moroccan Universities) with 18 attributes including SSG, CGPA, GBG, and NEB were selected	The proposed web tool presented successfully resulted in three profiles with slow, moderate, and fast learners
[18]	Profile and cluster students based on academic performance	K-Means - Weka	Student demographics and data on exam marks were retrieved through the Oman Education Portal database	Four student performance levels were identifies
[19]	Develop a framework to anticipate the learning behavior of education institutes by using best-suited data mining methods	Hierarchical, Nonhierarchical clustering algorithm.	Data of learning management system (Student Information System-(SIS), and Moodle data used by institutions.	Successfully gave a framework to assess students' performance in academics for decision-making.
[20]	Cluster students based on subject categories of computer core, mathematics, and general.	K-means - Weka	Data of undergraduate students from Mehran University, Pakistan analyzed based on academic data (sessional, mid, final exam)	Students perform better in computer-core subjects as compared to mathematics
[21]	Support programming learning using the MK-means algorithm	Modified K-means	Evaluation logs (problem-solving solution and test score) of Algorithm and Data Structures collected from the Online Judge platform	Derived statistical features and patterns that can be used to suggest improvement in the programming course
[22]	Develop practical guidance for clustering method to find student categories and characteristics	K-means	Real data set. Covers 26 departments of Mae Fah Luang (MFU) University, Ta-sud, Muang, Chiang Rai Thailand.	Students with good school-academic backgrounds do well in the first year.
[23]	Present an evaluation method that compares different methods for clustering using internal and external parameters	K-means, EM, Spectral, Agglomerative, DBSCAN, K-medoids	University of Tartu's, (Estonia), and Moodle data was used with 9 data sets in total having 15 features.	Agglomerative and K-medoid algorithms proved better with more than 10 features for normalized datasets

[24]	Explore the relationship between students' digital activities and their academic performance	Various Clustering approaches	Activity log data of 1st-semester undergraduate online course using Blackboard LMS at a University in Saudi Arabia 2019–2020	Students having different digital profiles can exhibit similar academic achievement.
[25]	Conduct detailed mining on the data of student performance.	K-means	Transcripts of three subjects named probability theory, mathematical statistics, and linear algebra, Shenyang Jianzhu University China.	141 students marked as 0 (scores above 80), 135 students marked as 1 (60 points), and 74 students marked as 2 (low scores in math and algebra).
[26]	Visualize the occupation or profession for the career of graduate students using a data mining approach.	K-means	141 graduate students' data of Computer Science Laurea degree University of Florence (Italy) with the period of batches from 2001-2008.	Results show that the majority of students followed the order mentioned in the ideal career.
[27]	Experiment clustering approach for EDM	Latent class, K-means, model base	Account data of 233 teachers divided into two categories: Novice (1 to 3 years of experience in teaching) and veteran (more than 3 years of experience).	Latent Class Analysis (LCA) is better than the K-means algorithm.
[28]	Prediction of evaluation by teachers based on the university student's dataset.	Various clustering algorithms	29 articles were selected after inclusion and exclusion criteria evaluation.	Research about teacher valuation with student input analysis using machine algorithms.
[29]	Detect the unusual behavior of teachers while analyzing his/her students	K-means	Simulation data from an Iraqi university system.	Student role concerning stage and semester and the role of lecturer has a significant impact on students' degree
[30]	Analyze student skills based on their problem-solving.	K-means clustering technique, Naïve Bayes	200 samples for training data of students with 9 attributes on computer science, mathematics, CSE aero science, and marine department programs	From the analysis of the results, many students got very good skill performance in problem-solving

[31]	Visualize online learner's activities to show the unrevealed facts or information	K-means	Data comprised 46524 records of the C programming course of MOOC at Kathmandu University of Nepal	After experimenting with an optimal number for clusters, k=5 was found to be the best
[32]	Develop relations that can help to make decisions based on the education variables nature.	Hierarchical clustering algorithm	The data used was from statistics from Yemen's official education sources.	20 closed rest regions found in the third cluster that show a level of education to the decision maker.
[33]	Develop an e-learning analytical approach method	K-means, Prediction analysis	An English e-learning course data of 120 students from the Pittsburgh Science of Learning Centre repository.	Teachers' decisions concerning the global answers of students are more efficient in a Virtual Learning Environment.
[34]	Highlight research in cluster-sampling and knowledge discovery databases	Clustering technique	The data set considered for this work is the population of China students being registered in primary and secondary education schools.	Proving by graphical representation that the number of students enrolling in secondary school after clearing primary school has increased in China
[35]	Develop a unified model with classification and clustering to enhance the employability prediction of students	Two-level clustering (K-means, K-star), random forest, random tree, simple cart	Considered information from students within the program of Engineering and Masters in Computer Applications from different universities (India)	Student's employability prediction is improved using the model with 96.78% accuracy and 0.937 Kappa value
[36]	Developing a trust model with data mining methods to be used in the current education system as a tool for strategy management	K-means	The dataset of the School of Computer Science, Engineering and Applications - Bharathidasan University, India.	The predicted pass and fail number of students in which pass students was 36, and 1 was a failed student while 1 was an absent student.

[37]	Study the application of political and ideological management of the college education system.	K-means	60 samples containing data for students from work record tables.	The clustering data mining model is verified to be successful with a percentage of 30%, 63%, and 7% of the total sample.
[38]	Provide comparison to check the efficiency of various techniques	Prediction and clustering	Not applicable	Gender, hometown, location, and family income is an overriding factor for student performance
[39]	Develop an analysis model for research on evaluation by teachers.	Improved K means clustering	Teacher evaluation data in which 200 questionnaires were present using a network evaluation system.	Improved clustering algorithm concluded to be best fit when the system is not perfect for teaching evaluation or not involving complete student participation.
[40]	Establish classification rules between the performance of students and the master program of those students and also cluster similar academic performances.	C45 and K-means	Closed-ended data of 25 questions were collected from 2nd and 3rd-year students at Information, Communication Technologies, Faculty of Natural Sciences University of Tirana, Albania	The best results were shown Parameter values as 20 for cluster, 10 for seed, and 1000 for maximum iteration
[41]	Present a clustering approach to anticipate interest groups of students.	K-means	Data collected from a survey from the author's organization consisting of a total of 176 data points	Communication Club has a higher accuracy of 93.22%
[42]	Analyze the assessment pattern of basic education and provide a mechanism to give strength by suggesting improvements.	K-means	Data of all 159 909 schools that were working in the years 2007, 2009, 2011, and 2013, different states in (Brazil).	It is concluded that reducing the number of schools in small clusters is best to identify the worst income reason.

[43]	Develop a system that can learn from the data itself.	Clustering	Not applicable.	Provided a review collectively of disparate entities in the educational sector.
[44]	Analysis of the use of clustering techniques.	Multiple clustering approaches	Not applicable.	Clustering techniques serve as an efficient first step to process data quickly

Observing the temporal view of the studies depicted in Figure 2, we can see a steady focus of research spanning between 2013 to 2023, with an increase in 2017, 2022, and 2023. Although researchers have focused on EDM before the COVID-19 pandemic, some researchers have attributed the increased research in EDM to this outbreak [9][10]. Many educational institutes had to switch to and rely on digital platforms. This also led to an additional influx in the educational data being generated, creating more opportunities to experiment and analyze educational content.

Focusing on the region of the experiments, we can observe from Figure 3 that research is not restricted to a particular region and spans various countries across the globe. However, a large chunk of the research has emerged from India and China. We can also observe from Figure 4 that research based on clustering in education has been slightly more focused on conference publications as compared to journal publications.

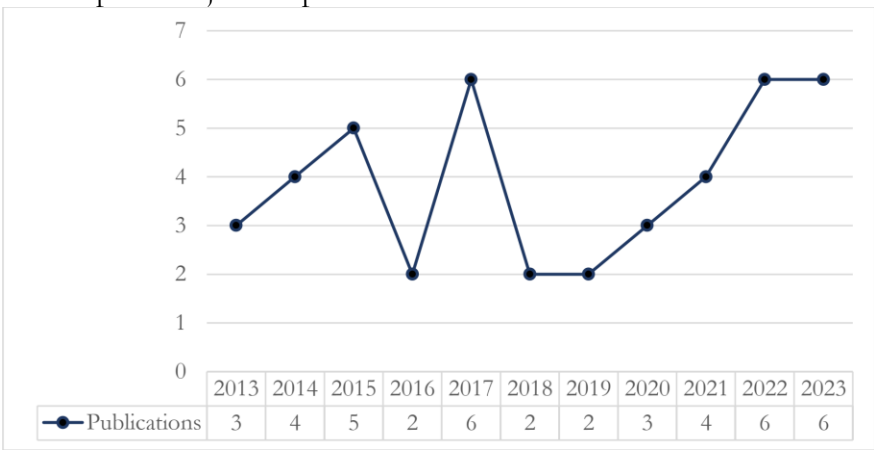

Figure 2. Temporal View of the Reviewed Literature

Discussion:

RQ1: What educational problems are being analyzed using cluster analysis?

From the literature review (Table 2), we can see that several educational problems have been targeted using a wide range of clustering methods in the reviewed literature, with extensive research being carried out on analyzing student performance in a course or a degree program [6][11][12][13][14][15][16][17]. Another educational problem targeted has been to cluster students based on the class of learners [18][19][20][21][22]. Research has also focused on clustering students based on their assignment submission patterns [3], activity in Moodle [1][19][23][16], use of Learning Management Systems [24], and finding patterns through their engagement in a course [10]. The emerging cluster patterns have been used to visualize and refine a learning environment [7], help provide interventions to reduce drop-outs [6], understand student competency in courses [25][15][20], and provide guidance pertaining to careers [26]. Cluster-based research has not only focused on finding patterns in student behavior but also on analyzing teacher patterns [27][28][29]. It is evident from the reviewed literature that the results

of cluster-based analysis can successfully be used to establish a framework to not only assess students' performance but also to shape pedagogy.

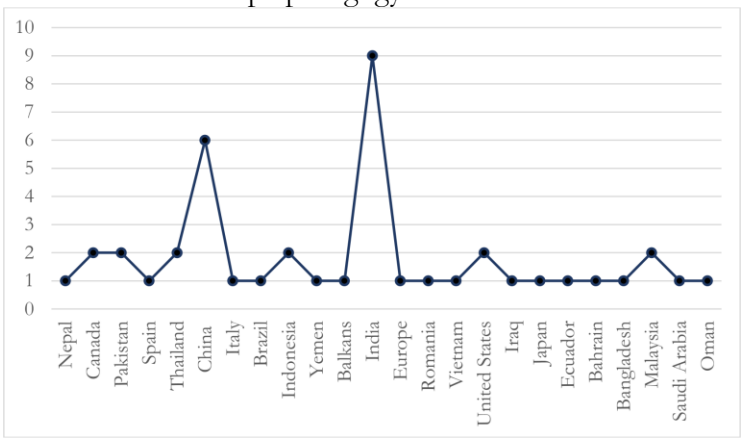

Figure 3. Countries Covered in the Review

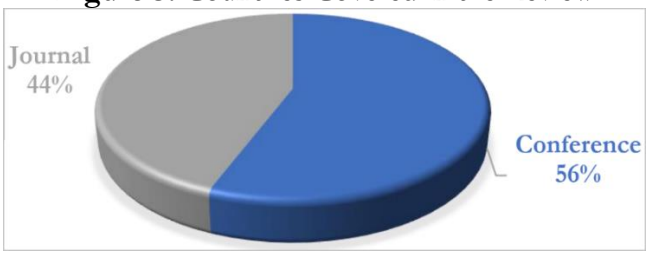

Figure 4. Article Coverage

RQ2: What are the techniques frequently selected for cluster analysis in EDM?

K-means clustering has been utilized extensively and far more than any other approach in the reviewed literature. Among the reviewed literature, a total of 30 of the 43 reviewed papers (69.76%) have used K-means clustering in their research. From analyzing student performance while submitting assignments [3], monitoring student performance to avoid drop-outs [6], clustering students based on their abilities to suggest future career options [26], clustering students based on their problem-solving abilities [30], clustering online learners' based on their activities [31], understanding learning styles [10], detecting teacher behavior while assessing students [29], to creating varied graduate profiles [17], and analyzing variance in student performance across different subject categories [20], K-means has been used to explore and tackle varied educational issues. Each of the other clustering approaches has been a focus of three of the reviewed studies with clustering focusing on hierarchical and non-hierarchical approaches being used for understanding student clusters in the subject of mathematics [7], better understanding the influence of educational variables [32], and understanding learning behavior [19]. EM clustering algorithm has been used at a more generalized level in the reviewed studies [1][23].

RQ3: Which parameters/attributes are considered for clustering-based educational research?

The last review question focused on the use of parameters in clustering-based EDM research. The most common parameters used in the reviewed papers have been summarized in Table 3. Parameters are an important consideration whenever carrying out EDM research. Performance-based features such as CPGA and internal assessment [1][2][20], assignment marks and marks in a semester [5][20], test scores, mid and final assessment [20][17], data obtained from online platforms comprising of time spent and interaction as well as feedback results [29] and many other features have been used in the reviewed papers. As we can observe from Table 3, the student learning data has been most commonly used in the reviewed literature, featuring in almost 44% of the literature. Data from Learning Management Systems has also been

extensively used to cluster students, followed by data collected through surveys and targeted questionnaires.

Table 3. Common Parameters and Algorithms used

Parameters	Ref.
Learning Data from a Learning Management System	[1][3][7][24][19][31][23][33][29][16][10]
GPA, Learning / Academic data	[2][3][6][34][35][18][11][36][26][24][30][12][37][13][14][15] [17][20][22]
Demographic, Behavior and Extra Features, General data	[22][4]
Online data: Exercise / Interactive / Learning	[3][7][21][10]
Logical, reasoning, numerical aptitude, personality, parents' qualification	[30][38]
Survey / Questionnaire	[39][40][41]

There are several approaches used in the field of EDM to analyze the performance of students across various domains. From the overall literature review, it is concluded that clustering is widely being utilized to analyze various facets of educational outcomes. Consistent with the findings of Dutt et al., several papers focused on exploring data emerging from e-learning platforms [5]. However, an exploration of data from intelligent tutoring systems still needs further exploration. Another similarity between Dutt et al. and Aulakh et al. was the use of K-means in a majority of the reviewed papers [5][9]. It was observed that clustering combined with other algorithms, especially classification (forming clusters first and then based on revealed characteristics performing further analysis), leads to tailored and better interventions [4][35][11][14].

Although several studies have been conducted to investigate and find patterns in student data, it is essential to employ the resultant findings for improving the educational infrastructure. Future directions in the field of cluster-based analysis can focus on the use of these techniques in conjunction with other data mining techniques such as classification and regression to tailor interventions and shape pedagogical policies. Integration of these approaches in e-learning systems with direct and instant feedback resulting in a dynamic e-learning environment can also be a possible research direction. Longitudinal studies can then be conducted to observe the long-term effects of interventions. There were certain limitations to this study including the limited number of knowledge sources explored for the review. This may have resulted in missing some research on cluster-based analysis that may have been published on some other knowledge source. This leaves room for future studies to explore a wider range of knowledge sources which will lead to a better understanding of the focus of cluster-based analysis in education.

Conclusion:

This paper aims to help researchers, educators, and policymakers better to understand the potential scope and benefits of using clustering within the domain of education by providing a comprehensive review of the current state of research by identifying common methods employed in educational contexts. The review explored a total of 43 papers extracted from five sources. Upon the review of the literature, it was found that various clustering algorithms have been explored by researchers. Although several papers have focused on the use of EM, X-means, and hierarchical clustering techniques, K-means was found to be the leading and most utilized clustering algorithm being the focus of almost 70% of papers. Clustering has been used to target several educational issues with the prime focus of research being the analysis of student performance in a course or a degree program followed by the analysis of the class of learners.

Student academic data has most commonly been explored in clustering-based research. Future directions in the field of cluster-based analysis can focus on comparing clustering techniques and the use of these techniques in conjunction with other data mining techniques such as classification, regression, and association rule mining to tailor interventions and shape pedagogical policies.

References:

[1] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for improving Educational process mining," ACM Int. Conf. Proceeding Ser., pp. 11–15, 2014, doi: 10.1145/2567574.2567604.

[2] S. Wijayanti, Azahari, and R. Andrea, "K-Means cluster analysis for students graduation (case study: STMIK widya cipta dharma)," ACM Int. Conf. Proceeding Ser., vol. Part F129684, pp. 20–23, Jun. 2017, doi: 10.1145/3108421.3108430.

[3] C. Geng, W. Xu, Y. Xu, B. Pientka, and X. Si, "Identifying Different Student Clusters in Functional Programming Assignments: From Quick Learners to Struggling Students," SIGCSE 2023 - Proc. 54th ACM Tech. Symp. Comput. Sci. Educ., vol. 1, pp. 750–756, Mar. 2023, doi: 10.1145/3545945.3569882.

[4] B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach," J. Med. Syst., vol. 43, no. 6, pp. 1–15, Jun. 2019, doi: 10.1007/S10916-019-1295-4/METRICS.

[5] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," IEEE Access, vol. 5, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[6] "'Importance of Data Mining in Higher Education System.'" Accessed: Dec. 19, 2023. [Online]. Available: https://www.researchgate.net/publication/269751502_Importance_of_Data_Mining_in_Higher_Education_System

[7] M. C. Desmarais and F. Lemieux, "Clustering and Visualizing Study State Sequences", Accessed: Dec. 19, 2023. [Online]. Available: https://www.educationaldatamining.org/EDM2013/papers/rn_paper_33.pdf

[8] B. Kitchenham, "Procedures for Performing Systematic Reviews," 2004, Accessed: Dec. 19, 2023. [Online]. Available: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf

[9] K. Aulakh, R. K. Roul, and M. Kaushal, "E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review," Int. J. Educ. Dev., vol. 101, p. 102814, Sep. 2023, doi: 10.1016/J.IJEDUDEV.2023.102814.

[10] P. Nuankaew, P. Nasa-Ngium, and W. S. Nuankaew, "Self-Regulated Learning Styles in Hybrid Learning Using Educational Data Mining Analysis," ICSEC 2022 - Int. Comput. Sci. Eng. Conf. 2022, pp. 208–212, 2022, doi: 10.1109/ICSEC56337.2022.10049322.

[11] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/J.COMPEDU.2017.05.007.

[12] Harwati, A. P. Alfiani, and F. A. Wulandari, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)," Agric. Agric. Sci. Procedia, vol. 3, pp. 173–177, Jan. 2015, doi: 10.1016/J.AASPRO.2015.01.034.

[13] L. Chen, M. Li, and Y. Chen, "Research on Course Score Analysis Based on K-Means Clustering Algorithm," Proc. - 2022 2nd Asia-Pacific Conf. Commun. Technol. Comput. Sci. ACCTCS 2022, pp. 485–488, 2022, doi: 10.1109/ACCTCS53867.2022.00104.

[14] M. Rahman, M. I. Ahmed, and M. S. Hossain, "Analysis of Student's Achievement through Educational Data Mining," Proc. - 2021 Int. Conf. Inf. Syst. Adv. Technol. ICISAT 2021, 2021, doi: 10.1109/ICISAT54145.2021.9678406.

[15] J. Liu, "The Application of K-Means Clustering Algorithm in the Quality Analysis of College English Teaching," Proc. - 2022 Int. Conf. Educ. Netw. Inf. Technol. ICENIT 2022, pp. 1–4, 2022, doi: 10.1109/ICENIT57306.2022.00009.

[16] M. Bucos and B. Dragulescu, "Student cluster analysis based on Moodle data and academic performance indicators," 2020 14th Int. Symp. Electron. Telecommun. ISETC 2020 - Conf.

Proc., Nov. 2020, doi: 10.1109/ISETC50328.2020.9301061.

[17] L. Najdi and B. Er-Raha, "Implementing cluster analysis tool for the identification of students typologies," Colloq. Inf. Sci. Technol. Cist, vol. 0, pp. 575–580, Jul. 2016, doi: 10.1109/CIST.2016.7804852.

[18] S. J. S. Alalawi, I. N. M. Shaharanee, and J. M. Jamil, "CLUSTERING STUDENT PERFORMANCE DATA USING k-MEANS ALGORITHMS," J. Comput. Innov. Anal., vol. 2, no. 1, pp. 41–55, Jan. 2023, doi: 10.32890/JCIA2023.2.1.3.

[19] M. A. Job and J. Pandey, "Academic Performance Analysis Framework for Higher Education by Applying Data Mining Techniques," ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 1145–1149, Jun. 2020, doi: 10.1109/ICRITO48877.2020.9197925.

[20] "Clusters of Success: Unpacking Academic Trends with K-Means Clustering in Education." Accessed: Dec. 19, 2023. [Online]. Available: https://www.researchgate.net/publication/375635449_Clusters_of_Success_Unpacking_Academic_Trends_with_K-Means_Clustering_in_Education

[21] M. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran, and K. Nakamura, "Educational Data Mining to Support Programming Learning Using Problem-Solving Data," IEEE Access, vol. 10, pp. 26186–26202, 2022, doi: 10.1109/ACCESS.2022.3157288.

[22] N. Iam-On and T. Boongoen, "Generating descriptive model for student dropout: a review of clustering approach," Human-centric Comput. Inf. Sci., vol. 7, no. 1, pp. 1–24, Dec. 2017, doi: 10.1186/S13673-016-0083-0/FIGURES/23.

[23] D. Hooshyar, Y. Yang, M. Pedaste, and Y. M. Huang, "Clustering Algorithms in an Educational Context: An Automatic Comparative Approach," IEEE Access, vol. 8, pp. 146994–147014, 2020, doi: 10.1109/ACCESS.2020.3014948.

[24] A. Bessadok, E. Abouzinadah, and O. Rabie, "Exploring students digital activities and performances through their activities logged in learning management system using educational data mining approach," Interact. Technol. Smart Educ., vol. 20, no. 1, pp. 58–72, Feb. 2023, doi: 10.1108/ITSE-08-2021-0148/FULL/XML.

[25] R. Gu, X. Jing, D. Zhao, L. Cai, and H. Gao, "Research and application of improved k-means algorithm based on educational big data," Proc. - 2022 Int. Conf. Comput. Eng. Artif. Intell. ICCEAI 2022, pp. 746–749, 2022, doi: 10.1109/ICCEAI55464.2022.00157.

[26] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," Expert Syst. Appl., vol. 42, no. 13, pp. 5508–5521, Aug. 2015, doi: 10.1016/J.ESWA.2015.02.052.

[27] B. Xu, M. Recker, X. Qi, N. Flann, and L. Ye, "Clustering Educational Digital Library Usage Data: A Comparison of Latent Class Analysis and K-Means Algorithms," J. Educ. Data Min., vol. 5, no. 2, pp. 38–68, Jul. 2013, doi: 10.5281/ZENODO.3554633.

[28] R. Ordoñez-Avila, N. Salgado Reyes, J. Meza, and S. Ventura, "Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review," Heliyon, vol. 9, no. 3, p. e13939, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13939.

[29] L. A. N. Muhammed, "Educational Data Mining: Analyzing Teacher Behavior based Student's Performance," Proc. - 2021 4th Int. Conf. Comput. Informatics Eng. IT-Based Digit. Ind. Innov. Welf. Soc. IC2IE 2021, pp. 181–185, 2021, doi: 10.1109/IC2IE53219.2021.9649366.

[30] M. Mayilvaganan and D. Kalpanadevi, "Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques," Procedia Comput. Sci., vol. 47, no. C, pp. 62–75, Jan. 2015, doi: 10.1016/J.PROCS.2015.03.184.

[31] S. Shrestha and M. Pokharel, "Machine Learning algorithm in educational data," Int. Conf. Artif. Intell. Transform. Bus. Soc. AITB 2019, Nov. 2019, doi: 10.1109/AITB48515.2019.8947443.

[32] F. Alqasemi, S. Al-Hagree, A. Aqlan, K. M. A. Alalayah, Z. Almotwakl, and M. Hadwan, "Education Data Mining for Yemen Regions Based on Hierarchical Clustering Analysis," 2021 Int. Conf. Technol. Sci. Adm. ICTSA 2021, Mar. 2021, doi: 10.1109/ICTSA52017.2021.9406544.

[33] A. M. De Morais, J. M. F. R. Araújo, and E. B. Costa, "Monitoring student performance using data clustering and predictive modelling," Proc. - Front. Educ. Conf. FIE, vol. 2015-February,

no. February, Feb. 2015, doi: 10.1109/FIE.2014.7044401.
[34] A. Abdulahi Hasan and H. Fang, "Data Mining in Education: Discussing Knowledge Discovery in Database (KDD) with Cluster Associative Study," ACM Int. Conf. Proceeding Ser., May 2021, doi: 10.1145/3469213.3471319.
[35] P. Thakar, A. Mehta, and Manisha, "A unified model of clustering and classification to improve students' employability prediction," Int. J. Intell. Syst. Appl., vol. 9, no. 9, pp. 10–18, Sep. 2017, doi: 10.5815/IJISA.2017.09.02.
[36] M. Durairaj and C. Vijitha, "Educational Data mining for Prediction of Student Performance Using Clustering Algorithms," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 4, pp. 5987–5991, 2014.
[37] L. Huang, "Teaching management data clustering analysis and implementation on ideological and political education of college students," Proc. - 2016 Int. Conf. Smart Grid Electr. Autom. ICSGEA 2016, pp. 308–311, Nov. 2016, doi: 10.1109/ICSGEA.2016.61.
[38] R. K. Rambola, M. Inamke, and S. Harne, "Literature review- techniques and algorithms used for various applications of educational data mining (EDM).," 2018 4th Int. Conf. Comput. Commun. Autom. ICCCA 2018, Dec. 2018, doi: 10.1109/CCAA.2018.8777556.
[39] P. Gu and Y. Zheng, "Cluster analysis on the teaching evaluation data from college students," 13th Int. Conf. Comput. Sci. Educ. ICCSE 2018, pp. 839–844, Sep. 2018, doi: 10.1109/ICCSE.2018.8468864.
[40] A. Ktona, D. Xhaja, and I. Ninka, "Extracting relationships between students' academic performance and their area of interest using data mining techniques," Proc. - 6th Int. Conf. Comput. Intell. Commun. Syst. Networks, CICSyN 2014, pp. 6–11, Mar. 2014, doi: 10.1109/CICSYN.2014.18.
[41] V. Bahel, S. Malewar, and A. Thomas, "Student Interest Group Prediction using Clustering Analysis: An EDM approach," Proc. 2nd IEEE Int. Conf. Comput. Intell. Knowl. Econ. ICCIKE 2021, pp. 481–484, Mar. 2021, doi: 10.1109/ICCIKE51210.2021.9410741.
[42] T. G. Ramos, J. C. F. Machado, and B. P. V. Cordeiro, "Primary Education Evaluation in Brazil Using Big Data and Cluster Analysis," Procedia Comput. Sci., vol. 55, pp. 1031–1039, Jan. 2015, doi: 10.1016/J.PROCS.2015.07.061.
[43] A. Dutt, "Clustering Algorithms Applied in Educational Data Mining," Int. J. Inf. Electron. Eng., 2015, doi: 10.7763/IJIEE.2015.V5.513.
[44] N. V. Krishna Rao, N. Mangathayaru, and M. Sreenivasa Rao, "Evolution and prediction of radical multi-dimensional e-learning system with cluster based data mining techniques," Proc. - Int. Conf. Trends Electron. Informatics, ICEI 2017, vol. 2018-January, pp. 701–707, Jul. 2017, doi: 10.1109/ICOEI.2017.8300793.