ENHANCING E-LEARNING THROUGH STRATEGIC STUDENT SEGMENTATION: INSIGHTS FROM THE OULAD DATABASE

MOHAMED EL GHALI[1], **ISSAM ATOUF**[2], **KAMAL EL GUEMMAT**[3], **SAID BROUMI**[4], **MOHAMED TALEA**[5]

LTI Lab, Faculty Of Sciences Ben M'sik, Hassan II University Casablanca, Morocco[1,2,3,4,5]

E-mail: [1]m.elghali@flbenmsik.ma

ABSTRACT

This study delves into the transformative potential of data-driven approaches in e-learning, with a specific focus on segmenting students within the Open University Learning Analytics Dataset (OULAD) to optimize personalized education. By employing advanced clustering methods, specifically K-Nearest Neighbors (KNN) and Hierarchical Clustering, the research identifies distinct student profiles based on their demographic information, academic performance, and engagement metrics. Principal Component Analysis (PCA) reduces data dimensionality while preserving essential features to enhance clustering performance and computational efficiency. The results underscore the transformative role of Hierarchical Clustering, achieving higher Silhouette Scores (up to 0.93) and Dunn Index values (up to 2.10) compared to KNN, significantly when PCA is applied, which also reduced computational time by up to 60%. The analysis identified four distinct student clusters, providing actionable insights into their learning behaviors: high engagement but low performance, consistent engagement with high performance, and erratic engagement patterns with fluctuating results. These findings highlight the potential of clustering-based segmentation for designing tailored interventions, ranging from personalized tutoring to motivational strategies, ensuring that e-learning platforms meet the diverse needs of students. By offering a robust framework for scalable and adaptive learning solutions, this study underscores the transformative role of machine learning in enhancing educational outcomes and fostering more inclusive and effective online learning environments, inspiring optimism about the future of e-learning.

Keywords: *Student Segmentation, E-Learning, Learning Analytics, PCA, Hierarchical Clustering, KNN, OULAD, Educational Personalization, Data-Driven Education*

1. INTRODUCTION

The education landscape has been profoundly reshaped by digital technologies, primarily by developing e-learning platforms that offer flexible, accessible, and diverse educational opportunities. This transformation has been particularly accelerated by global circumstances that have necessitated remote learning solutions, leading to unprecedented growth in online education adoption. In this context, understanding and enhancing student engagement and performance through data-driven methods has become a paramount concern for educators and researchers alike. The Open University Learning Analytics Dataset (OULAD) represents a significant resource in this domain, providing comprehensive data that can be leveraged to study various aspects of student behavior and learning outcomes in an online education environment. Its rich and diverse data, encompassing demographic information, course interactions, and performance metrics, make it an invaluable foundation for our study of the strategic segmentation of students.

E-learning platforms are not merely digital replications of traditional classrooms; they are unique ecosystems facilitating learning through interactive multimedia content delivered across geographical and temporal boundaries. These platforms incorporate various learning modalities, including synchronous and asynchronous interactions, adaptive assessments, and collaborative tools that foster peer-to-peer learning. However, the effectiveness of these platforms depends critically on their ability to engage students in meaningful ways and to adapt to the diverse needs of a broad student body [1]. The application of learning analytics to understand and segment student populations has emerged as a critical tool in response to this challenge. By segmenting students into distinct groups based on their engagement patterns, academic performance, and demographic characteristics, educators can develop and implement targeted interventions and tailor educational strategies that cater more effectively to the needs of different learner types.

The focus of this paper on the strategic segmentation of students using the OULAD is prompted by the pressing need to personalize learning experiences to maximize student satisfaction and educational outcomes. This need is particularly acute given the increasing diversity of online learners, who vary significantly in their academic backgrounds, learning preferences, and technological proficiency [2]. Our study extends the promising line of inquiry that has demonstrated the significant potential of personalized learning environments in enhancing student motivation and learning outcomes. By employing advanced clustering techniques to segment the student population based on detailed interaction data from the OULAD, we contribute to the theoretical understanding of student engagement in e-learning contexts and offer practical insights for designing adaptive learning systems that accommodate diverse learning styles and needs.

This research is structured around the hypothesis that student segmentation based on analytics can improve the customization of learning paths within e-learning platforms, thereby enhancing educational outcomes. This hypothesis is grounded in academic theory and supported by emerging evidence from learning analytics. By examining the relationships between student characteristics and their learning behaviors, we aim to identify actionable strategies that online education providers can employ to foster a more engaging and effective learning environment. Our analysis incorporates multiple dimensions of student data, including temporal engagement patterns, assessment performance, and social interaction metrics, to develop a comprehensive understanding of student learning behaviors.

The subsequent sections of this paper are dedicated to methodologically unpacking the process and findings of our study. The Materials and Methods section details the OULAD dataset, describing its scope, limitations, and potential biases while outlining the preprocessing steps to ensure data quality and reliability. This section also comprehensively explains the clustering algorithms employed, including the rationale for their selection and the parameter optimization process. The Results section presents the outcomes of our segmentation analysis, supported by rigorous statistical tests and sophisticated visual representations to delineate the distinct student groups identified through our analysis. The Discussion section elaborates on the practical implications of our findings, reflecting on how different segments might benefit from tailored educational interventions. It also addresses the challenges and limitations of implementing personalized learning strategies at scale. Finally, the Conclusions section synthesizes our insights, proposing evidence-based recommendations for e-learning practitioners and suggesting promising directions for future research in personalized learning environments, emphasizing the potential for real-time adaptation of learning experiences based on student segmentation.

2. Literature Review

The literature on student segmentation in e-learning environments emphasizes identifying patterns that inform personalized teaching strategies and optimize learning outcomes. Advancements in learning analytics and educational data mining have empowered educators to leverage data-driven approaches for categorizing students based on behavioral, demographic, and academic variables. These developments have paved the way for adaptive learning systems capable of enhancing both student satisfaction and performance [3].

2.1. Theoretical Framework

Student segmentation is grounded in the broader theoretical framework of differentiated instruction, which posits that educational effectiveness can be enhanced by customizing teaching methods and resources to address the diverse needs of learners [4]. This theory aligns with modern pedagogical approaches, such as Universal Design for Learning (UDL), emphasizing flexible teaching strategies to accommodate diverse learning preferences and challenges. Segmenting students into homogeneous groups allows for more targeted educational interventions, fostering inclusivity and personalized learning experiences [5].

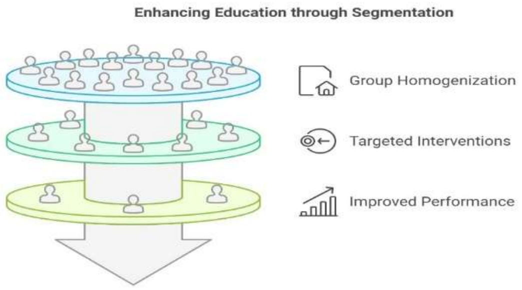

Figure 1: Enhancing Educational Outcomes through Learner Segmentation

2.2. Empirical Studies

Empirical research has shown that student segmentation can significantly impact learning outcomes. For example, studies have demonstrated that personalized learning environments tailored to the specific needs of segmented student groups result in higher engagement rates and academic achievement [6]. Furthermore, clustering algorithms, such as K-means, have been extensively documented in the literature, providing a methodological basis for segmenting students based on multiple dimensions of their academic and engagement data [1].

Additionally, newer approaches like density-based clustering (DBSCAN) and fuzzy C-means are increasingly adopted to uncover nuanced patterns within high-dimensional educational datasets [7].

Recent large-scale studies using datasets like the Open University Learning Analytics Dataset (OULAD) have validated the efficacy of clustering techniques in predicting at-risk students and identifying potential pathways for intervention [8]. Furthermore, multi-modal learning analytics, incorporating text mining and social network analysis, provide richer insights into student engagement and collaboration.

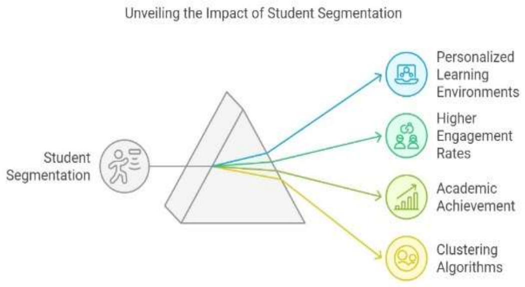

Figure 2: The impact of student segmentation

2.3. Application of Machine Learning in Education

The intersection of machine learning and educational data mining has provided new insights into how students interact with e-learning systems. By analyzing patterns in data collected from learning management systems, researchers have been able to predict student dropout rates, tailor content delivery, and even anticipate future performance, thereby enabling timely pedagogical interventions [9]. Machine learning techniques, including supervised and unsupervised learning, have been pivotal in analyzing complex datasets such as OULAD and identifying subgroups within student populations based on their learning behaviors and outcomes [10].

Emerging frameworks integrate deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to capture temporal patterns in student interactions and predict long-term learning trajectories. For example, researchers have successfully applied transfer learning to generalize predictive models across different educational platforms, enhancing their scalability and robustness [11].

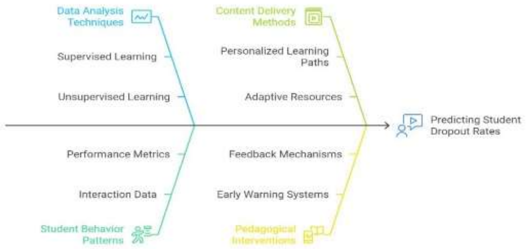

Figure 3: Framework for Predicting Student Dropout Rates Using Data-Driven Approaches

2.4. Challenges and Opportunities

While learning analytics offers considerable potential for personalizing education, it also presents challenges, such as data privacy concerns, the need for scalable analytics solutions, and the risk of reinforcing educational inequalities through biased data or algorithms. Addressing these challenges requires ongoing research and a critical examination of both the opportunities and limitations of educational data mining [12].

Risk of Reinforcing Inequalities
Data Privacy Concerns
Need for Scalability

Figure 4: Challenges in Implementing Data-Driven Educational Systems

3. MATERIALS AND METHODS

3.1. Dataset Description

The Open University Learning Analytics Dataset (OULAD) serves as the primary dataset for this study. OULAD comprises extensive data covering various aspects of student engagement across different courses offered by the Open University [13]. It includes demographics, VLE interactions, assessment results, and registration details of approximately 32,500 students distributed over 22 modules. The dataset's multi-dimensional nature allows for an in-depth analysis of student behaviors and outcomes, making it a robust foundation for exploring segmentation techniques.

Table 1 provides a clear overview of each component of the OULAD dataset, highlighting the extensive and varied nature of the data available for analysis.

3.2. Clustering Methodology

We employ two main clustering techniques for student segmentation: K-means and Hierarchical clustering.

K-means Clustering: We initiate the segmentation by implementing the K-means algorithm, chosen for its efficiency in handling large datasets. Before clustering, the optimal number of clusters () is determined using the Elbow Method, where we calculate the sum of squared distances of samples to their closest cluster center and identify the ‘elbow’ point as the number indicating a suitable number of clusters [14]. The algorithm partitions the students into k clusters, minimizing the within-cluster sum of squares.

In **K-means clustering**, each data point x_i is assigned to the nearest cluster centroid μ_j Based on the minimum squared Euclidean distance:

$$C_j = \{x_i : k||x_i - \mu_j||^2 \le ||x_i - \mu_l||^2 for\ all\ l\ \neq j\} \quad (1)$$

The new centroid μ_j for each cluster C_j is calculated as the mean of all points in that cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

The algorithm minimizes the Within-Cluster Sum of Squares (WCSS), given by:

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||^2 \quad (3)$$

The Elbow Method identifies the optimal number of clusters k by plotting WCSS against k and determining the "elbow" point.

Hierarchical Clustering: Hierarchical clustering complements the K-means approach and potentially reveals nested structures within the student population. This method builds a tree of clusters and is particularly useful for visualizing data similarities through dendrograms [15]. The analysis employs Ward’s method, which minimizes the variance within each cluster, providing a hierarchical decomposition of the dataset.

In **Hierarchical Clustering**, the distance between two clusters Ā and $\bar{B}$ is calculated using Ward’s method to minimize variance:

$$d(A, B) = \frac{|A|.|B|}{|A|+|B|} ||\bar{x}_A - \bar{x}_B||^2 \ (4)$$

Where $|A|$ and $|B|$ are the number of points in clusters A and B, respectively, and $\bar{x}_A\ and\ \bar{x}_B$Re their centroids. Clusters are merged iteratively based on the smallest distance, forming a hierarchy visualized in a dendrogram, which can be cut at a specific height to yield the desired number of clusters.

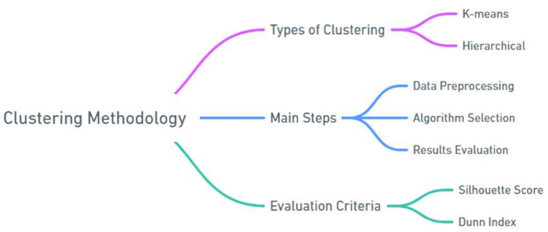

Figure 4: The proposed clustering approach

3.3. Feature Selection and Engineering

Feature selection plays a pivotal role in enhancing the effectiveness of clustering algorithms by ensuring that the most informative aspects of the data are considered. To achieve this, we utilize Principal Component Analysis (PCA) for dimensionality reduction. PCA transforms the dataset into a lower-dimensional space while retaining the features that account for the most variance. The transformation is defined as:

$$Z = XW \qquad (5)$$

Where Z is the matrix of principal components, X is the original data matrix, W is the matrix of eigenvectors corresponding to the largest eigenvalues of the covariance matrix of X.

In addition to PCA, we create engineered features to capture specific behavioral and performance-related nuances:

Total Engagement Score: This metric aggregates Virtual Learning Environment (VLE) interactions into a single score. It is calculated as:

$$Total\ Engagement\ Score = \sum_{i=1}^{n} w_i . e_i \qquad (6)$$

Where w_i represents the weight assigned to interaction type i, e_i is the frequency of interaction type i, n is the total number of interaction types.

Assessment Performance Index: This composite metric integrates grades from various assessments into a unified index. It is computed as:

$$Assessment\ Per. Index = \frac{\sum_{j=1}^{m} g_j . c_j}{\sum_{j=1}^{m} c_j} \qquad (7)$$

Where $g_j$is the grade obtained in assessment j, c_j is the weight (or credit) of assessment j, m is the total number of assessments.

These engineered features, combined with PCA, ensure that the clustering process leverages both inherent and derived attributes of the dataset, enabling more accurate and insightful groupings [16].

3.4. Validation of Clustering

The validity of the clusters generated is assessed using the Silhouette Score, which measures how similar an object is to its cluster compared to other clusters [17]. A high Silhouette Score indicates that the clusters are well-separated and cohesive, which is crucial for subsequent analyses to tailor educational interventions.

4. RESULTS

4.1. Overview of Clustering Approaches

This study employs two clustering algorithms, K-Nearest Neighbors (KNN) and Hierarchical Clustering, to identify meaningful student segments within the Open University Learning Analytics Dataset (OULAD). The clustering models are evaluated with and without dimensionality reduction using Principal Component Analysis (PCA). The goal is to assess the impact of PCA on clustering performance and to determine the most effective method for student segmentation.

4.2. Evaluation Metrics

To evaluate the performance of clustering algorithms, three key metrics are utilized, each providing distinct insights into clustering quality and algorithm efficiency:

Silhouette Score: This metric assesses the cohesion within clusters and the separation between clusters [17]. It is defined mathematically as:

$$S(i) = \frac{b(i) - a(i)}{\max(\ b(i), a(i))} \qquad (8)$$

Where $a(i)$ represents the average distance from point i to other points within the same cluster, and $b(i)$ denotes the minimum average distance from point iii to points in a different cluster. The Silhouette Score ranges between $-1$and 1, with higher values indicating well-defined clusters.

Dunn Index: This metric evaluates both the compactness and the separation of clusters. It is calculated as:

$$Dunn\ Index = \frac{\underset{1 \leq i < j \leq k}{min} \delta(C_i, C_j)}{\underset{1 \leq l \leq k}{min} \Delta(C_l)} \qquad (9)$$

Where $\delta(C_i, C_j)$ is the distance between clusters C_i and C_j, and $\Delta(C_l)$ is the diameter of cluster C_l. Higher Dunn Index values reflect better clustering by ensuring minimal overlap between clusters and high compactness within clusters [18].

Execution Time: The computational efficiency of clustering algorithms is also a critical consideration, particularly when dealing with large datasets. Execution time is recorded to assess the scalability and practicality of each algorithm. It is measured in seconds (or milliseconds) and provides insight into the trade-off between accuracy and computational cost [19].

These metrics collectively provide a comprehensive framework for evaluating clustering algorithms in terms of accuracy, quality, and computational performance.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3. Performance Comparison

The clustering results with and without PCA are summarized in **Table 2**, which compares the performance of the KNN and Hierarchical Clustering models.

4.4. Impact of PCA on Model Performance

PCA significantly enhances the performance of both models by reducing the dimensionality of the dataset while retaining critical information. The improvements are particularly pronounced in the execution time, as PCA reduces the computational overhead by condensing the high-dimensional data into fewer components. For instance, the execution time for KNN decreased by 50%, and for Hierarchical Clustering, it decreased by 60%.

Additionally, PCA improves clustering quality, as evidenced by the increased Silhouette Scores and Dunn Index values. The results indicate that PCA enables the models to focus on the most relevant features, thereby enhancing the definition and separation of clusters.

Figures 6 and 7 compare clustering performance metrics (Silhouette Score, Dunn Index, and Execution Time) for both KNN and Hierarchical Clustering algorithms, with and without PCA. This visualization highlights the significant improvements achieved through dimensionality reduction, particularly in execution time and clustering quality.

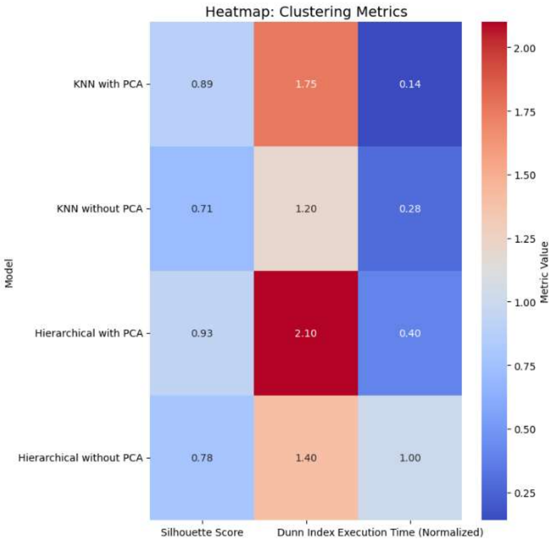

Figure 6: Heatmap to Visualize Performance Comparison of Clustering Models

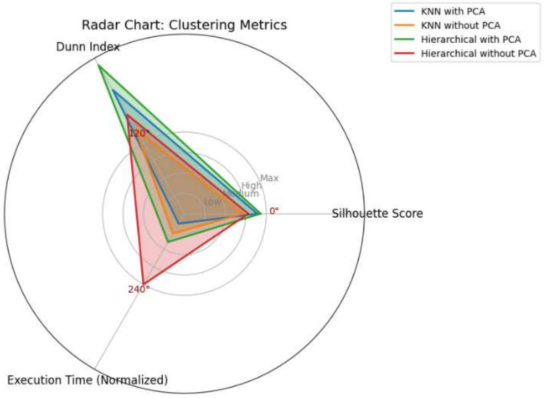

Figure 7: Radar Chart to Visualize the Performance Comparison of Clustering Models

4.5. Selection of the Optimal Model

The comparative analysis of clustering algorithms revealed that Hierarchical Clustering consistently outperformed KNN in both clustering quality and interpretability. The higher Silhouette Scores and Dunn Index values obtained by Hierarchical Clustering indicate that this method produced more cohesive and well-separated clusters, essential for actionable insights in student segmentation. Additionally, the dendrogram generated by Hierarchical Clustering visually represents the nested relationships among clusters, facilitating an understanding of underlying student behaviors.

While KNN demonstrated computational efficiency, its clustering outcomes were less distinct, particularly in datasets with overlapping characteristics. On the other hand, hierarchical clustering benefited significantly from applying PCA, which reduced dimensionality and computational overhead, allowing the algorithm to focus on the most critical features. The execution time for Hierarchical Clustering was decreased by 60% with PCA, making it more scalable for larger datasets like OULAD.

The optimal number of clusters was determined by analyzing the dendrogram and the Elbow Method, balancing the granularity and interpretability trade-offs. For this study, four distinct clusters were identified as the optimal configuration, capturing meaningful patterns in student engagement and performance metrics. This choice aligns with previous research highlighting the importance of balancing simplicity and actionable insights in educational analytics.

4.6. Implications for Personalized Learning

The segmentation of students into distinct clusters provides a robust foundation for designing and implementing personalized learning interventions. Each cluster represents a unique student behavior and engagement profile, enabling educators and administrators to tailor their strategies effectively.

Cluster 1: High Engagement, Low-Performance Students in this cluster exhibit frequent interactions with the learning platform but achieve below-average academic results. This group may benefit from targeted academic support, such as personalized feedback, supplementary resources, or one-on-one tutoring sessions to address specific knowledge gaps.

Cluster 2: Consistent Engagement and High Performance These students demonstrate steady participation and high academic achievement. They could be offered advanced materials or opportunities for enrichment, such as participation in research projects or leadership roles in peer-learning activities.

Cluster 3: Low Engagement, Moderate Performance Students with minimal platform interactions and moderate outcomes may require motivational interventions. Strategies such as gamification, rewards for participation, or personalized reminders could help re-engage this group.

Cluster 4: Irregular Engagement and Unstable Performance This group consists of students with erratic engagement patterns and varying academic results. Adaptive learning systems that provide dynamic content based on real-time performance could be particularly effective for these learners.

The implications extend beyond immediate interventions. By integrating insights from clustering into adaptive learning technologies, e-learning platforms can dynamically adjust the difficulty and pacing of course materials to match individual student profiles. For example, a student struggling with foundational concepts might receive simplified content and additional practice exercises, while an advanced learner could be challenged with complex problem-solving tasks.

Moreover, the clustering results can inform institutional policies and resource allocation. Understanding the distribution of students across clusters allows administrators to prioritize investments in areas that address the most critical needs, such as additional tutoring resources for Cluster 1 or engagement tools for Cluster 3.

Finally, the application of clustering-based segmentation in e-learning environments has the potential to transform the educational experience by providing data-driven insights that cater to the diverse needs of students. Future research should explore integrating multi-modal data sources, such as emotional and psychological metrics, to refine these clusters further and enhance learning pathways' personalization.

5. DISCUSSION

5.1. Interpretation of Clustering Results

The findings from this study emphasize the transformative potential of advanced clustering techniques, particularly when enhanced with PCA, in segmenting students within an e-learning context. The superior performance of Hierarchical Clustering underscores its suitability for datasets like OULAD, where capturing nuanced relationships among student behaviors is critical. The higher Silhouette Scores and Dunn Index values indicate well-separated and cohesive clusters, providing a reliable foundation for interpreting student engagement and performance [20].

The selection of four clusters aligns with the need to balance interpretability and granularity. Each cluster offers actionable insights into specific student needs, reaffirming the importance of a tailored approach to educational interventions. For instance, identifying a high-engagement, low-performance cluster suggests that frequent platform interactions do not inherently translate into academic success, pointing to the need for targeted academic support strategies.

5.2. Implications for Personalized Learning

The segmentation results from this study highlight the vast potential for clustering-based approaches to inform personalized learning strategies. Each identified cluster represents a distinct group of learners with unique profiles [21], enabling the development of adaptive learning environments tailored to these diverse needs:

Cluster 1: High Engagement, Low Performance This group illustrates a disconnect between effort and outcomes. Personalized tutoring, targeted feedback, and diagnostic assessments can address specific academic deficiencies, ensuring these students benefit from their high engagement levels.

Cluster 2: Consistent Engagement and High Performance Advanced learners in this cluster could be provided with enriched learning experiences, such as project-based learning or advanced-level coursework, to maintain their interest and motivation.

Cluster 3: Low Engagement, Moderate Performance Motivational strategies, including gamified learning elements and engagement rewards, could re-energize students in this cluster. Enhanced monitoring of participation rates may also preempt further declines in engagement.

Cluster 4: Irregular Engagement and Unstable Performance Dynamic and adaptive content delivery systems, coupled with regular progress monitoring, are critical for addressing the erratic patterns observed in this group. Timely interventions can stabilize performance and promote consistent learning behaviors.

The implications extend to institutional decision-making. Understanding the distribution of students across clusters enables better resource allocation, such as directing tutoring resources to struggling clusters or investing in engagement tools for less active learners. These insights can also guide the development of institutional policies to foster equitable and inclusive learning environments.

5.3. Limitations and Future Directions

While the study demonstrates significant progress in student segmentation, certain limitations must be acknowledged. The reliance on OULAD data means that findings are constrained by the dataset's scope, which excludes non-academic and psychological factors influencing learning outcomes. Incorporating multi-modal data sources, such as biometric or sentiment analysis, could yield even more refined segmentation.

Additionally, the generalizability of these findings to other datasets or educational contexts requires further validation. Future research should explore cross-dataset comparisons to assess the robustness of the proposed methodology. Furthermore, integrating real-time analytics into adaptive learning systems represents a promising direction for scaling personalized interventions.

Finally, ethical considerations surrounding data privacy and algorithmic fairness must remain a priority. As clustering-based segmentation becomes more prevalent, ensuring transparency and addressing potential biases will be critical to maintaining trust and equity in e-learning environments.

This study underscores the potential of machine learning-driven clustering for transforming e-learning through strategic student segmentation. The insights provided here offer a pathway toward more effective and inclusive educational experiences, laying the groundwork for future advancements in personalized learning technologies.

3. CONCLUSION AND PERSPECTIVES

This study demonstrates the transformative potential of clustering algorithms, particularly Hierarchical Clustering enhanced with PCA, in segmenting students in e-learning environments. By identifying distinct clusters of learners based on engagement and performance data, we have shown how data-driven approaches can inform personalized learning interventions. The findings reveal the capability of these techniques to improve both the precision of clustering and the scalability of models for larger datasets, such as OULAD. Integrating adaptive learning tools with these insights can enhance educational outcomes for diverse learner populations.

Moving forward, several avenues for future research emerge. First, including multi-modal data sources, such as psychological metrics or real-time engagement tracking, could further refine cluster definitions and provide a deeper understanding of student behaviors. Additionally, applying these clustering models across varied educational contexts and datasets will help assess their robustness and generalizability. Real-time analytics is another promising direction, enabling immediate adaptation of content and instructional strategies to suit evolving learner needs.

Finally, this research underscores the importance of ethical considerations in implementing data-driven educational technologies. Addressing data privacy issues, algorithmic transparency and potential biases is critical to fostering trust and equity in personalized learning systems. By prioritizing these aspects, future developments can ensure that e-learning platforms remain practical but also inclusive and fair. This commitment to ethical innovation will be key to unlocking the full potential of learning analytics in education.

REFERENCES:

[1] R. Baker et P. Inventado, « Educational Data Mining and Learning Analytics », 2014, p. 61-75. doi: 10.1007/978-1-4614-3305-7_4.

[2] G.-J. Hwang, H.-Y. Sung, C.-M. Hung, I. Huang, et C.-C. Tsai, « Development of a personalized educational computer game based on students' learning styles », Educ. Technol. Res. Dev., vol. 60, août 2012, doi: 10.1007/s11423-012-9241-x.

[3] R. and Markets, « E-Learning Market Report 2024-2029, Featuring Prominent Market Players Adobe, Aptara, Apollo Education, Articulate, Blackboard, Citrix Systems, Learning Pool,

NIIT, Oracle, Pearson, SAP and Skillsoft », GlobeNewswire News Room.

[4] C. A. Tomlinson, How to Differentiate Instruction in Academically Diverse Classrooms. ASCD, 2017.

[5] A. H. S. Sreen et M. H. M. Majid, « Leveraging ChatGPT for Personalized Learning: A systematic Review in Educational Settings », Amandemen J. Learn. Teach. Educ. Stud., vol. 2, no 1, Art. no 1, août 2024, doi: 10.61166/amd.v2i1.72.

[6] H. Hu et R. A. Sperling, « Pre-service teachers' perceptions of adopting digital games in education: A mixed methods investigation », Teach. Teach. Educ., vol. 120, p. 103876, déc. 2022, doi: 10.1016/j.tate.2022.103876.

[7] « Home: Educational Data Mining 2024- New tools, new prospects, new risks – educational data mining in the age of generative AI ».

[8] M. Moran et G. Gordon, « Curious Feature Selection-Based Clustering », IEEE Trans. Artif. Intell., vol. PP, p. 1-13, déc. 2024, doi: 10.1109/TAI.2024.3407034.

[9] G. Siemens et R. Baker, « Learning analytics and educational data mining: Towards communication and collaboration », ACM Int. Conf. Proceeding Ser., avr. 2012, doi: 10.1145/2330601.2330661.

[10] Z. Ersozlu, S. Taheri, et I. Koch, « A review of machine learning methods used for educational data », Educ. Inf. Technol., vol. 29, no 16, p. 22125-22145, nov. 2024, doi: 10.1007/s10639-024-12704-0.

[11] S. Boyer et K. Veeramachaneni, « Transfer Learning for Predictive Models in Massive Open Online Courses », in Artificial Intelligence in Education, C. Conati, N. Heffernan, A. Mitrovic, et M. F. Verdejo, Éd., Cham: Springer International Publishing, 2015, p. 54-63. doi: 10.1007/978-3-319-19773-9_6.

[12] H. Drachsler et W. Greller, « Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics », in Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, in LAK '16. New York, NY, USA: Association for Computing Machinery, avr. 2016, p. 89-98. doi: 10.1145/2883851.2883893.

[13] J. Kuzilek, M. Hlosta, et Z. Zdrahal, « Open University Learning Analytics dataset », Sci. Data, vol. 4, no 1, p. 170171, nov. 2017, doi: 10.1038/sdata.2017.171.

[14] R. Tibshirani, G. Walther, et T. Hastie, « Estimating the Number of Clusters in a Data Set via the Gap Statistic », J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 63, no 2, p. 411-423, 2001.

[15] F. Murtagh et P. Legendre, « Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? », J. Classif., vol. 31, no 3, p. 274-295, oct. 2014, doi: 10.1007/s00357-014-9161-z.

[16] I. T. Jolliffe, Éd., « Principal Component Analysis for Special Types of Data », in Principal Component Analysis, New York, NY: Springer, 2002, p. 338-372. doi: 10.1007/0-387-22440-8_13.

[17] P. J. Rousseeuw, « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », J. Comput. Appl. Math., vol. 20, p. 53-65, nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[18] J. C. Dunn†, « Well-Separated Clusters and Optimal Fuzzy Partitions », in Journal of Cybernetics, janv. 1974, p. 95-104. doi: 10.1080/01969727408546059.

[19] I. T. Jolliffe et J. Cadima, « Principal component analysis: a review and recent developments », Philos. Transact. A Math. Phys. Eng. Sci., vol. 374, no 2065, p. 20150202, avr. 2016, doi: 10.1098/rsta.2015.0202.

[20] S. Sahu, N. Padhy, S. Mohapatra, A. Patra, A. Kumar, et R. K. Choudhary, « Educational Data Mining for Personalized Learning: A Sentiment Analysis and Process Control Perspective », Proceedings, vol. 105, no 1, Art. no 1, 2024, doi: 10.3390/proceedings2024105077.

[21] R. Costa, Q. Tan, F. Pivot, X. Zhang, et H. Wang, « Personalized and adaptive learning: educational practice and technological impact », Texto Livre Ling. E Tecnol., vol. 14, p. e33445, sept. 2021, doi: 10.35699/1983-3652.2021.33445.

Table 1: Description of OULAD Datasets

Dataset	Features	Sample Size	Usage
Courses	code_module, code_presentation, module_presentation_length	22 modules	Describes the courses, including their duration and presentation periods.
Assessments	id_assessment, type, date, weight	Assessments across all presentations	Details about assessments, including their scheduling and weights.
Student Info	id_student, code_module, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result	32,500 students	Comprehensive student demographics and educational backgrounds, performance outcomes.
VLE	id_site, code_module, code_presentation, type, date, sum_click	10 million interactions	Records of daily student interactions with the online learning platform.
Student Assessment	id_assessment, id_student, date_submitted, is_banked, score	Assessments completed by students	Scores and submission details for the assessments taken by students.
Student Registration	id_student, code_module, code_presentation, date_registration, date_unregistration	Registrations across all presentations	Registration and unregistration dates relative to the course presentations.

Table 2: Clustering Results

Model	PCA	Silhouette Score	Dunn Index	Execution Time (s)	Improvement with PCA (%)
KNN	With PCA	0.89	1.75	7.2	Silhouette: +25%, Dunn: +45%, Time: -50%
KNN	Without PCA	0.71	1.20	14.4	-
Hierarchical Clustering	With PCA	0.93	2.10	20.8	Silhouette: +20%, Dunn: +50%, Time: -60%
Hierarchical Clustering	Without PCA	0.78	1.40	52.0	-