

Unlocking Learner Engagement and Performance: A Multidimensional Approach to Mapping Learners to Learning Cohorts

Abstract

Learning analytics, located at the intersection of learning science, data science, and computer science, aims to leverage educational data to enhance teaching and learning. However, as educational data increases, distilling meaningful insights presents challenges, particularly concerning individual learner differences. This work introduces a comprehensive approach for designing and automatically mapping learners into meaningful cohorts based on diverse learning behaviors. It defines four critical contexts, including engagement, direction, repetitiveness, and orderliness, and generates practical learning cohorts from their combinations. The approach employs a time-series-based clustering method with K-means clustering using dynamic time warping to identify similar learning patterns. Statistical techniques like the Mann-Kendall test and Theil-Sen estimator further refine the process. A case study on data science courses validates the approach, offering novel insights into learner behavior. The contributions include a novel time series approach to characterizing learning behavior, new learning cohorts based on critical contexts, and a systematic method for automated cohort identification.

Keywords: learning analytics, cohort analysis, learning behaviors, clustering, time series

1 Introduction

Learning analytics is a dynamic and promising field that acts as a bridge between the realms of learning science, data science, and computer science (Clow, 2013; Ferguson, 2012; Siemens, 2013; Viberg et al., 2018). It operates at the intersection of these disciplines, aiming to harness the abundance of data generated within educational environments (Siemens, 2013). Based on this integration, learning analytics desires to propel the improvement of teaching and learning processes (García-Peñalvo, 2020), promote academic progress and performance (Wright et al., 2014), and build an optimized learner-oriented educational environment (Sedrakyan et al., 2020).

The primary objective of learning analytics is to extract valuable insights from the vast reservoir of educational data (Viberg et al., 2018). To this end, a wide range of data sources is involved, such as learning behaviors, usage of materials, and outcomes

of assessments (Samuelsen et al., 2019). In addition, to transform these raw data into actionable knowledge, the discipline leverages a manifold of techniques spanning data science and computational methodologies, including machine learning, natural language processing, process mining, and others (Viberg et al., 2018). Such techniques are used for unveiling hidden learning patterns, designing and adapting curricula, predicting academic performance, and developing optimized learning processes.

So far, many efforts have proved that these techniques are effective in achieving the core of learning analytics. Nevertheless, as the volume of educational data continues to expand at an unprecedented pace, distilling meaningful insights from this vast expanse poses a significant challenge. One of the challenges is to deal with important differences between individual learners.

Cohort analysis is a type of behavioral analytics that focuses on tracking and analyzing the behaviors or characteristics of specific groups of individuals (Pope, 1972). As a versatile and valuable tool, there are various domains where it ascertains beneficial. For example, in healthcare and clinical research, cohort studies are pivotal in identifying risk factors with disease development, assessing the effectiveness of medical treatments, and understanding the long-term outcomes of patients (Vincent et al., 2009). The learning analytics discipline also employs cohort analysis to achieve the intricate results of groups of learners with similar characteristics or experiences (Ferguson & Clow, 2015; Khalil & Ebner, 2017).

However, constructing meaningful cohorts is not trivial, requiring various contexts to be considered, including social or cultural, personal, and behavioral contexts. The social or cultural context includes cultural, linguistic background, and social interaction, while the personal context consists of demographics, learning goals, and prior knowledge. The behavioral context refers to how learners interact with course materials and navigate their learning journeys. Notably, existing research often focuses on learners' engagement patterns within behavioral aspects. For example, consider (Ferguson & Clow, 2015), which divides learners into explicit categories based on their engagement: samplers, strong starters, returners, mid-way dropouts, nearly there, late completers, and keen completers. While engagement patterns are a significant component of behavioral patterns, other valuable elements, such as regularity or sequence of activities, within behavioral data might not be fully utilized for constructing cohorts in traditional educational settings.

To support the automated construction of such meaningful cohorts, in this work, we propose a comprehensive approach for (i) designing cohorts from multiple aspects of learning behaviors and (ii) mapping learners into cohorts automatically from data. First, our approach defines multiple cohorts that effectively capture diverse learning behaviors. To achieve this, we consider four critical contexts: engagement, direction, repetitiveness, and orderliness. The engagement context centers on the length of learning records, while the direction indicates how learners interact with pre-determined guidelines over time. Also, the repetitiveness and orderliness contexts refer to the revisit and randomness of learning records, respectively. Then, based on the combinations of these four contexts, we define nine practical learning cohorts to conduct various analyses. Second, we suggest a systematic method for automatically mapping learners into pre-designed cohorts, focusing on the four contexts. It first applies the

time-series-based clustering approach to build clusters of similar learning records. To this end, we employ the K-means clustering (Krishna & Murty, 1999) with a dynamic time warping method (Müller, 2007). Then, we present a systematic approach to subdivide learning patterns within each dimension based on the barycenter of clusters. In this process, we also leverage statistical techniques, such as the Mann-Kendall test (Kendall, 1948; Mann, 1945), the Theil-Sen estimator (Sen, 1968; Theil, 1950), and the Pettitt test (Pettitt, 1979).

To validate our approach, we have collected learning data from three courses pertinent to the data sciences field. This dataset serves as the foundation for testing and refining our proposed approach. By applying our approach to these courses, we aim to demonstrate its effectiveness in uncovering meaningful insights using analytics techniques.

In summary, this paper has the following contributions:

- We propose new learning cohorts based on four critical contexts that provide a novel perspective on categorizing learners. This can enhance the understanding of learner behaviors and lead to more effective investigations and interventions.
- We suggest a systematic method for automatically mapping learners to the cohort groups using time-series-based clustering and statistical approaches. This method reduces the need for manual identifications of the suggested learning cohorts.
- We have evaluated the effectiveness and utility of the proposed approach in a real-life online education platform.

The rest of this paper is structured as follows. [Section 2](#) discusses the existing works related to our approach. In [Section 3](#), we present the four contexts for learning behaviors and define meaningful cohorts, while a series of steps for mapping learners into the cohorts is introduced in [Section 4](#). [Section 5](#) delivers the results of the application of our method. [Section 6](#) discusses the significance and the limitations of our approach. Finally, [Section 7](#) concludes the paper by summarizing our work.

2 Related work

The concept of learning cohorts in learning analytics revolves around the identification and analysis of subpopulations within a more extensive educational context (Deng et al., 2020; Kizilcec et al., 2013). The goal is to provide deeper knowledge of the learning process, cater to learners’ diverse needs, and improve personalized learning strategies. Applications of such cohorts extend into areas such as adaptive learning environments (El-Sabagh, 2021; Kabudi et al., 2021), where learning paths are personalized based on cohort characteristics; early identification of students at risk of dropping out, allowing for timely interventions; and enhancing group collaboration by assembling learners with complementary skills and learning preferences.

Constructing meaningful cohorts requires various contexts to be considered, including social or cultural, personal, and behavioral contexts. The social or cultural context encompasses not only learners’ cultural and linguistic backgrounds but also their socioeconomic status, which can influence access to technology and resources and their integration within the learning community (King & Trinidad, 2021). Factors such as

ethnicity, gender, and age also play critical roles in shaping the learning experience and outcomes (Dong & Kang, 2022). On the other hand, the personal context includes not only demographics and learning goals but also learners' prior educational background, personal interests, and life responsibilities (Hosen et al., 2021). These factors can significantly affect learners' motivation, time management, and overall engagement with the course material.

The behavioral context in learning analytics refers to how learners interact with course materials and navigate their learning journeys. It can be classified into *engagement*, *direction*, *repetitiveness*, and *orderliness* of learning behaviors. Engagement encompasses various aspects of how students interact with course content, including access frequency, discussion participation, and material utilization (Caspari-Sadeghi, 2022). Direction examines if learners follow prescribed learning paths (Wong et al., 2019). Repetitiveness reflects on how often students revisit materials, indicating areas of interest or difficulty (Sennhenn-Kirchner et al., 2018). Orderliness measures the consistency of engagement, with erratic patterns suggesting non-linear learning approaches (Cao et al., 2018).

Literature on learning cohorts predominantly focuses on the behavioral context and, more specifically, on engagement. Several frameworks have been proposed to categorize learners into cohorts using engagement as a key factor. Ferguson and Clow (2015) divides learners into explicit categories based on their engagement: samplers, strong starters, returners, mid-way dropouts, nearly there, late completers, and keen completers. Li and Baker (2018) categorized learners into groups such as disengagers, auditors, quiz-takers, and all-rounders, revealing how different engagement patterns, like watching lectures or participating in forums, influence course grades and completion. Moreover, Fredricks et al. (2004) provided a comprehensive framework for understanding student engagement, breaking it down into behavioral, emotional, and cognitive components. Henrie et al. (2015) explored these dimensions of engagement in the context of online courses, offering insights into how different forms of engagement affect student learning outcomes.

Data-driven approaches have been instrumental in categorizing learners into distinct groups based on their engagement in online courses. Wen et al. (2014) employed computational linguistic models to scrutinize Massive Open Online Course (MOOC) forum posts. Their focus was on understanding student motivation and cognitive engagement. They detected patterns indicative of at-risk students by conducting fine-grained content analyses of student interactions across three MOOCs. Building on this, Ramesh et al. (2014) advanced the understanding of student engagement in MOOCs by designing a framework using student behavioral cues to model engagement. By creating latent representations of student engagement types, their research not only aided in predicting course completions but also served as a tool for the early identification of students needing additional support. Additionally, Hussain et al. (2018) study utilizing machine learning to assess student engagement in e-learning at the Open University identified low-engagement students, underscoring the efficacy of data-driven approaches in enhancing student performance and engagement in virtual learning environments.

Among various machine learning algorithms, clustering algorithms have been extensively utilized to categorize learners into distinct groups based on their interaction with learning materials and platforms. K -means clustering, for instance, has been frequently applied to segment students by examining engagement levels and learning outcomes, proving effective in identifying distinct learning patterns, such as high achievers, at-risk students, or irregular learners, thus facilitating targeted educational interventions (Abdallah Moubayed & Lutfiyya, 2020; Bharara et al., 2018; Labrović et al., 2023). Additionally, hierarchical clustering provides a more detailed perspective of student engagement, uncovering complex groupings rooted in similar learning behaviors or preferences (Antonenko et al., 2012; Deng et al., 2020).

Few approaches have leveraged time series analysis to explore engagement patterns. Kizilcec et al. (2013) applied time-series analysis to examine students’ engagement patterns in MOOCs, identifying distinct trajectories and providing insights into how learners’ engagement with course materials changes throughout a course. This dynamic view of learning highlights patterns like consistent engagement or late surges in activity, often linked with outcomes such as course completion or dropout. Furthermore, Matcha et al. (2019) utilized sequence mining and temporal analytics to track and categorize learner activities and their temporal sequences in online learning environments, revealing patterns such as intensive early engagement followed by a decline, and late engagement spikes, labeling these as early starters and last-minute learners.

While engagement is a significant component of behavioral context, existing approaches may not fully leverage other valuable elements, such as direction, repetitiveness, and orderliness, for constructing cohorts. This oversight highlights the limitations of current frameworks, underscoring the need for a more comprehensive framework to understand and categorize learner behaviors effectively. In response to these limitations, this work considers the three additional dimensions to construct more refined and homogeneous cohorts of learners.

Furthermore, we present a novel approach that combines clustering and time series analysis by employing time series clustering techniques to derive various learning cohorts. This method allows for identifying distinct temporal patterns over time, considering not only engagement, but also direction, repetitiveness, and orderliness. This approach facilitates a more dynamic categorization of learners based on their interactions with the course over its duration. It offers a new perspective on the categorization of learners, enhancing the potential for personalized and adaptive learning strategies.

3 A Taxonomy of Cohorts for Learning Behaviors

This section introduces a taxonomy to build cohorts for learning behaviors. To this end, we first introduce the concept of learning behaviors with a novel presentation of them in [Subsection 3.1](#). Next, we present four dimensions for segmenting cohorts, i.e., engagement, direction, repetitiveness, and orderliness, in [Subsection 3.2](#). Lastly, we define cohorts that hold comparable learning patterns in [Subsection 3.3](#).

3.1 Learning Behaviors

A learning behavior refers to how students progress through a course in logical time. Logical time refers to the sequence of events, such as the first, second, third, etc., rather than the actual time.

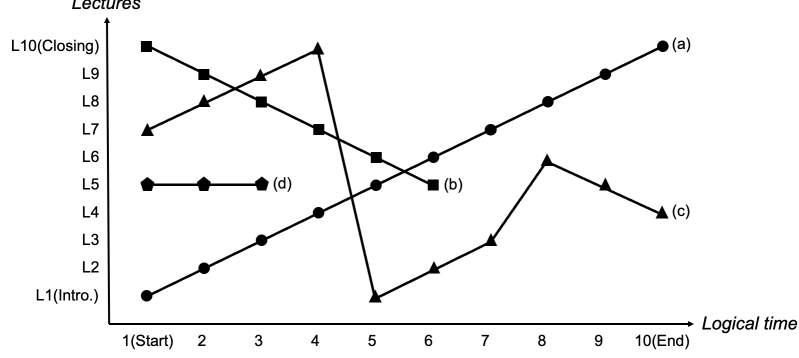


Figure 1: Examples of learning behaviors

Figure 1 visualizes learning behaviors of four learners in a logical time. The x-axis of the figure represents logical time, which means the order of learning events. On the other hand, the y-axis represents the learning activities within a course, especially lectures in this example. These activities typically follow a pre-defined sequence. For example, a course comprises several chapters, and even within a single chapter, the order of lectures is pre-determined. That is, the bottom part of the y-axis maps to the introductory activity in the course, while the top part maps to the closing activity.

For example, the behavior depicted in Figure 1(a) starts from the introduction lecture, i.e., L1, and advances ascendingly without any fluctuation, ending with the closing lecture, i.e., L10. This indicates a consistent and structured learning behavior, where the learner follows the pre-determined guideline from start to finish. Figure 1(b) has the learning sequences of starting from the closing lecture, proceeding in reverse order, but not arriving at the introduction lecture, i.e., midway drop out. This represents a reverse trend in the learner’s engagement with the course material. Figure 1(c) starts from the seventh lecture and ends with the fourth lecture. Unlike the previous patterns, they fluctuate as they jump back and forth between lectures. Therefore, this pattern is less structured to learning, and the learners engaged in this pattern may have explored specific topics or revisited certain sections. Lastly, Figure 1(d) only focuses on a single lecture, i.e., L5, with a recurrent exploration.

3.2 Dimensions to build cohorts for learning behaviors

Building cohorts for learning behaviors requires a structured assessment of learning behaviors based on various dimensions. In the following, we introduce dimensions that are used to build more refined and homogeneous cohorts of learners.

3.2.1 Engagement of learning behaviors

The first dimension in the proposed concept, i.e., engagement, is the length of lectures or activities within a course. As explained earlier, dividing learners into multiple cohorts or clusters is pivotal in existing cohort-based learning analytics field research. It indicates how thoroughly a learner has engaged in a course. That is, it enables classifying learners into various patterns based on their progress. Following the approach by (Ferguson & Clow, 2015), learners can be categorized into the following patterns: sampler, early dropout, midway dropout, completer, and others. In addition, we can divide the learners into observers (cf. Figure 1(b), (d)) or engagers (cf. Figure 1(a), (c)) based on their engagement.

3.2.2 Direction of learning behaviors

The second element for building cohorts is the direction of learning records. It refers to how learners interact with the course materials over time. More in specific, it identifies whether learners tend to follow the reference learning behaviors by the lecturer in a sequential manner (cf. Figure 1(a)), reverse the order of lectures (cf. Figure 1(b)), or display indifference towards the reference (cf. Figure 1(c), (d)). Here, sequential learners who follow the pre-determined reference hold an *ascending* trend. On the other hand, the learners who take the course materials in the opposite order represent a *descending* trend. Note that some learning behaviors cannot be categorized as ascending or descending because of the low number of lectures or high fluctuations.

3.2.3 Repetitiveness of learning behaviors

The third context for defining learning cohorts is repetitiveness, which refers to how learners revisit and engage with the same learning materials multiple times throughout their learning journey. This helps distinguish learners who explore a high degree of review and practice. In Figure 1, only pattern (d) holds the recurrent learning, while pattern (a)-(c) take each lecture only once. As depicted in the Figure 1(d), this factor results in a neutral or gentle slope.

3.2.4 Orderliness of learning behaviors

The last factor is the orderliness of learning records, which identifies and categorizes learners based on the regularity and consistency of the records. A higher randomness or fluctuation indicates that learners hold the lectures or activities more erratic and non-linearly, moving back and forth between different chapters or sections of the course content (cf. Figure 1(c)). Conversely, lower randomness occurs when learners follow the learning materials more orderly or evenly, progressing through the course content step by step, chapter by chapter(cf. Figure 1(a), (b)).

3.3 Context-based cohorts of learning behaviors

This section introduces 9 distinct learning cohorts derived from the four dimensions. Note that these cohorts are established through a methodical process that considers

all potential scenarios resulting from the various combinations of values within these dimensions.

After extensive deliberation, we ultimately established 9 coherent cohorts. Table 1 presents the identified learning cohorts, categorized according to four key dimensions. The subsequent portion of this section provides a comprehensive breakdown of each of these cohorts with visual representation for clarity as depicted in Figure 2.

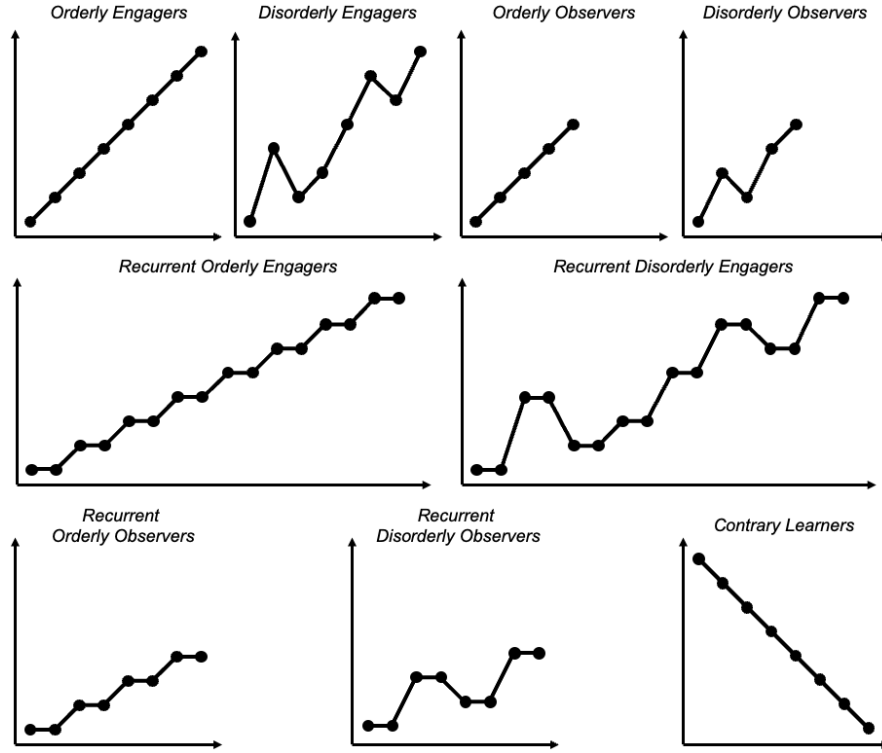


Figure 2: A graphical presentation of learning behaviors for each cohort

3.3.1 Orderly engagers

Orderly engagers represent a learning cohort of learners who successfully complete the course by strictly adhering to the provided guidelines without deviating into selective exploratory behaviors. These individuals consistently adhere to the prescribed learning path and avoid deviations from the instructional framework. As such, Figure 2 shows that the slope of this group exhibits a nearly 1.0 incline, underscoring their consistent adherence to the course's structure.

Table 1: The context-based learning cohorts considering four dimensions

Engaging	Engagement Observing	Direction		Repetitiveness		Orderliness		Cohorts
		Forward	Backward	Recurrent	Non-recurrent	Orderly	Disorderly	
✓		✓			✓	✓		Orderly Engagers
✓		✓			✓		✓	Disorderly Engagers
	✓	✓			✓	✓		Orderly Observers
	✓	✓			✓		✓	Disorderly Observers
✓		✓		✓		✓		Recurrent Orderly Engagers
✓		✓		✓			✓	Recurrent Disorderly Engagers
	✓	✓		✓		✓		Recurrent Orderly Observers
	✓	✓		✓			✓	Recurrent Disorderly Observers
✓		✓	✓	✓	✓	✓	✓	Contrary Learners

3.3.2 Disorderly engagers

Disorderly engagers primarily align with the reference and sufficiently engage the course, yet they have a tendency for selective explorations within certain course materials, disregarding the reference sequence. This dynamic behavior introduces noticeable fluctuations in their engagement pattern, while the overall slope depicts a nearly 1.0 incline, as indicated in [Figure 2](#).

3.3.3 Orderly observers

Similar to the orderly engagers, *orderly observers* strictly adhere to the course guidelines without engaging in selective explorations. However, this group discontinues their participation in the course either during the early stages or at a mid-point. Despite their initial alignment with the course structure, these learners discontinue their engagement before reaching the course's conclusion. This distinctive behavior is readily discernible in [Figure 2](#).

3.3.4 Disorderly observers

Disorderly observers closely resemble the disorderly engagers in their behaviors. However, the distinction lies in their inability to complete the entire course materials. These learners undertake selective explorations within specific course materials but discontinue their participation prematurely, either during the early stages or at a mid-point. This unique behavior is also revealed in [Figure 2](#). Also, a subset of this cohort can intentionally bypass specific course materials, thus deviating from the prescribed guidelines and exhibiting incomplete engagement with the course. In that case, this behavior results in a steep slope with an incline exceeding 1.0.

3.3.5 Recurrent orderly engagers

Recurrent orderly engagers closely follow the guidelines, reflecting a determined learning approach similar to the orderly engagers. However, their distinctive trait lies in their detail-oriented learning pace. These learners consistently revisit the same course materials multiple times, aiming for a thorough understanding of the content. Therefore, their pursuit of a clear understanding drives this intentional repetition. As such, their learning pattern exhibits an incline that remains below 1.0, as depicted in [Figure 2](#).

3.3.6 Recurrent disorderly engagers

Recurrent disorderly engagers also exhibit a shared trait of purposeful repetition in their learning methodology. However, their behaviors sometimes diverge from the guidelines as they engage in selective explorations within the course materials. In contrast to strictly complying with the prescribed path, these learners intentionally revisit certain sections while exploring others. This combination of repetition and selectiveness characterizes their unique learning pattern. As [Figure 2](#) indicates, this pattern is characterized by a gentle slope coupled with significant fluctuations.

3.3.7 Recurrent orderly observers

Recurrent orderly observers exhibit behavior similar to that of the recurrent orderly engagers. However, a key distinction emerges in their course completion. While both groups share the trait of deliberate repetition in their learning process, this cohort deviates by discontinuing their course participation, either during the early stages or at a mid-point. We can readily observe this behavior in [Figure 2](#). Also, a subset of this cohort can exclusively engage with course materials with similar content within the same chapter. In this case, this distinctive behavior contributes to a flat-shaped engagement pattern.

3.3.8 Recurrent disorderly observers

Recurrent disorderly observers have the behaviors that exhibit deliberate repetition alongside selective explorations within the course materials. However, they discontinue their course participation prematurely as observers. We can readily observe this behavior in [Figure 2](#). Also, a subset of this cohort can follow the behaviors of the repeated exploration of a couple of specific course materials across various chapters.

3.3.9 Contrary learners

Contrary Learners are a collection of individuals who study in a manner contrary to the provided course guidelines. Note that such patterns are not commonly observed in reality, as most students tend to follow the given guidelines or adopt a self-regulated approach to learning. This unique learning pattern results in a negative slope, as [Figure 2](#) depicts.

4 A method for mapping learners into learning cohorts

In [Section 3](#), we defined learning cohorts based on the four fundamental dimensions that underlie the behavior of learners: Orderly Engagers, Disorderly Engagers, Orderly Observers, Disorderly Observers, Recurrent Orderly Engagers, Recurrent Disorderly Engagers, Recurrent Orderly Observers, Recurrent Disorderly Observers, and Contrary Learners. This section aims to present a comprehensive method that provides a structured approach to map learners into distinct cohorts effectively. [Figure 3](#) describes an overview of the proposed approach. Initially, we start with time-series learning records of students and divide them into segments based on the length of their learning behaviors, such as segment 1, segment 2, etc ([Subsection 4.1](#)). Subsequently, for each of these segments, we employ time-series clustering techniques to create distinct clusters of learners ([Subsection 4.2](#)). Then, utilizing the barycenter (i.e., centroid) of these clusters, we assign them to specific learning cohorts ([Subsection 4.3](#)). This process involves three key steps, each corresponding to assessing one of the dimensions outlined in [Section 3](#). These dimensions include evaluating engagement, assessing direction and repetitiveness, and measuring orderliness. By analyzing how each cluster performs in these dimensions, we can effectively categorize it into one of the nine learning cohorts.

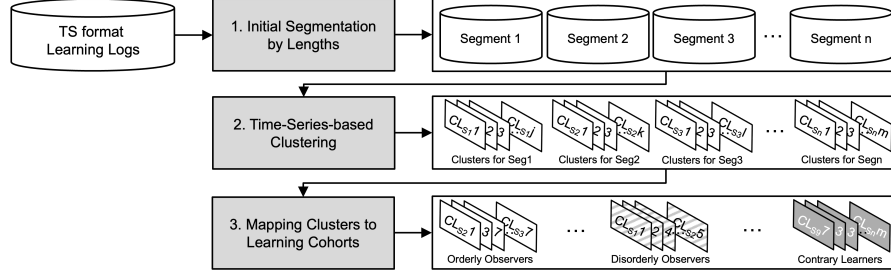


Figure 3: An overview of mapping learners into learning cohorts

4.1 Input data

We first explain the requisite input data for our proposed method. Typically, learning management systems maintain learning event logs, which constitute a collection of learning activities performed by individuals. Table 2 presents a fragment of a learning log.

Table 2: A fragment of a learning log

Event ID	Case ID	Activity	Timestamp	Logical Time	Pass/Fail	...
e1	c1	Lecture 1	2023-01-01 09:05:00	1	Pass	...
e2	c1	Lecture 2	2023-01-01 12:13:00	2	Pass	...
e3	c1	Lecture 3	2023-01-08 15:20:00	3	Pass	...
e4	c1	Lecture 4	2023-01-08 17:31:00	4	Pass	...
e5	c1	Lecture 5	2023-01-15 08:17:00	5	Pass	...
e6	c1	Lecture 6	2023-01-15 13:42:00	6	Pass	...
e7	c1	Lecture 7	2023-01-22 11:16:00	7	Pass	...
e8	c1	Lecture 8	2023-01-22 13:22:00	8	Pass	...
e9	c1	Lecture 9	2023-01-29 09:16:00	9	Pass	...
e10	c1	Lecture 10	2023-01-29 18:35:00	10	Pass	...
e11	c2	Lecture 1	2023-01-02 16:49:00	1	Fail	...
e12	c2	Lecture 9	2023-01-02 19:17:00	2	Fail	...
e13	c2	Lecture 2	2023-01-03 07:20:00	3	Fail	...
...	...	...	...	...	...	...
e67	c10	Lecture 2	2023-02-19 09:55:00	8	Pass	...

A learning event log encompasses numerous learning events, which serve as elemental units signifying a specific facet of learning behavior. These events have corresponding cases, which pertain to the learners involved. Additionally, details regarding the specific activity and the timestamp indicating when the event occurred are also logged. Through the temporal sequence of events within a given case, we can establish logical time—a chronological depiction of event order. Furthermore, these logs can involve other event and case attributes, such as pass or fail outcomes.

Based on this event log, we can construct a time-series-based representation describing the learning behaviors exhibited by the cases, as introduced in Figure 1. In this visualization, the x-axis indicates the chronological progression of events denoted as logical time, whereas the y-axis represents the established lecture sequence, ordered

by the pedagogical guidelines. Assuming the guideline follows a sequential numeric schema, beginning from Lecture 1 and ending at Lecture 10, the y-axis is systematically configured to span the numerical spectrum from 1 to 10 in a sequential order. For example, the event records of case c1 in Table 2 is transformed into a time-series dataset that incrementally ascends continuously: $\mathbf{x}_{c1} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. Likewise, the time-series data of the case c2 is represented as $\mathbf{x}_{c2} = (1, 9, 2)$.

By transforming learning logs into time-series data, we generate a time-series dataset $\mathbf{S} = \{\mathbf{x}_{c1}, \mathbf{x}_{c2}, \dots, \mathbf{x}_{c10}\}$ that encapsulates the learning behaviors of all learners. Figure 4(a) illustrates the learning behaviors of 10 cases.

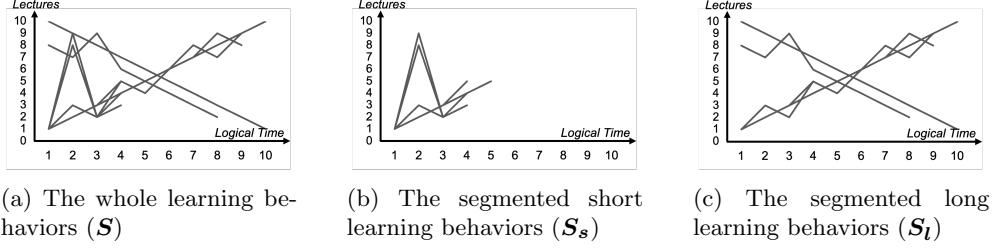


Figure 4: An example of time-series-format learning logs

Next, intending to build clusters with similar learning behaviors, we first segment learners based on the number of lengths of learning records. Opting for a user-by-user clustering approach rather than aggregating all instances into a single bucket enables us to perform time series clustering effectively. It ensures that we maintain a degree of homogeneity in the underlying data, enhancing the accuracy of subsequent clustering analyses. Also, the length of records aligns seamlessly with our goal of constructing meaningful learning cohorts since it belongs to one of the key dimensions of explaining learners' behavioral tendencies.

In our running example, we can divide our dataset into two segments based on a threshold length of 5. One segment $\mathbf{S}_s = \{\mathbf{x}_{c2}, \mathbf{x}_{c4}, \mathbf{x}_{c6}, \mathbf{x}_{c8}, \mathbf{x}_{c10}\}$ includes cases with learning records of equal to or less than 5 in length, while the other segment $\mathbf{S}_l = \{\mathbf{x}_{c1}, \mathbf{x}_{c3}, \mathbf{x}_{c5}, \mathbf{x}_{c7}, \mathbf{x}_{c9}\}$ comprises cases with records longer than 5. Figure 4(b) and (c) describe the segmented data with short and long behaviors, respectively.

4.2 Clustering of learning records

After the initial segmentation, we apply a clustering technique to each length group. Here, we employ a time-series-based clustering approach (Aghabozorgi et al., 2015), which makes clusters according to the hidden patterns within time-series data. Time-series clustering refers to the unsupervised partitioning of time-series data so that homogenous time-series are grouped based on a specific similarity measure. It aims to identify similarities within sequences, allowing for grouping similar sequences together.

In this method, we take the inputs of learning records presented in the form of time series data, as introduced in the previous section. That is, it encapsulates the sequential engagement of learners with course materials. Also, to enhance the effectiveness of clustering and to better capture intricate patterns, such as selective exploration behaviors, we leverage Dynamic Time Warping (DTW) (Müller, 2007). DTW enables us to measure the similarity between time series sequences by accommodating variations in the timing and pacing of time series data. It aims to find the optimal alignment between two sequences of data points by minimizing the cost. Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_m)$ be the sequences with the length of n and m , and $dist$ be a distance between sequences, e.g., Euclidean distance (Danielsson, 1980) or Pearson correlation (Cohen et al., 2009). Then, the cost is recursively measured with the following formula

$$DTW(\mathbf{x}, \mathbf{y}) = \min_{\pi} \sqrt{\sum_{(i,j) \in \pi} dist(x_i, y_j)^2} \quad (1)$$

, where $\pi = (\pi_1, \dots, \pi_K)$ is a path with a list of index pairs $\pi_k = (i_k, j_k)$ with $0 < i_{k-1} \leq i_k \leq i_{k-1} + 1 \leq n$ and $0 < j_{k-1} \leq j_k \leq j_{k-1} + 1 \leq m$. In such a step, dynamic programming and backtracking methods are utilized to derive the optimal alignment. Compared to traditional approaches that measure distances by assuming fixed-length sequences and synchronized data points, this technique is practical and valuable since it is designed to handle sequences with different lengths and accommodate time shifts.

Based on segmented data $\mathbf{S}_s$ and $\mathbf{S}_l$ in our running example, we can observe the process of dynamic time warping more concretely. Given two sequences, $\mathbf{x}_{c2} = (1, 9, 2)$ and $\mathbf{x}_{c4} = (1, 8, 2, 3)$, DTW maps these two sequences as follows: $((1, 1), (9, 8), (2, 2), (2, 3))$, resulting in the DTW distance of 1.41. On the other hand, for $\mathbf{x}_{c2} = (1, 9, 2)$ and $\mathbf{x}_{c6} = (1, 2, 3, 4, 5)$, the mapping generates these pairs: $((1, 1), (1, 2), (1, 3), (9, 4), (2, 5))$, leading to a distance of 6.24. Thus, even though the lengths of the vectors differ, we can compute distances using DTW, and the computed DTW distances among learning behaviors in $\mathbf{S}_s$ and $\mathbf{S}_l$ are presented in Table 3(a) and (b), respectively.

	$\mathbf{x}_{c2}$	$\mathbf{x}_{c4}$	$\mathbf{x}_{c6}$	$\mathbf{x}_{c8}$	$\mathbf{x}_{c10}$
$\mathbf{x}_{c2}$	0.00	1.41	6.24	6.24	5.48
$\mathbf{x}_{c4}$	1.41	0.00	5.39	5.39	5.20
$\mathbf{x}_{c6}$	6.24	5.39	0.00	1.41	1.00
$\mathbf{x}_{c8}$	6.24	5.39	1.41	0.00	1.41
$\mathbf{x}_{c10}$	5.48	5.20	1.00	1.41	0.00

(a) The DTW distances of learning sequences over cases in $\mathbf{S}_s$

	$\mathbf{x}_{c1}$	$\mathbf{x}_{c3}$	$\mathbf{x}_{c5}$	$\mathbf{x}_{c7}$	$\mathbf{x}_{c9}$
$\mathbf{x}_{c1}$	0.00	2.65	2.65	18.17	15.13
$\mathbf{x}_{c3}$	2.65	0.00	2.00	16.61	13.56
$\mathbf{x}_{c5}$	2.65	2.00	0.00	16.67	13.78
$\mathbf{x}_{c7}$	18.17	16.61	16.67	0.00	3.00
$\mathbf{x}_{c9}$	15.13	13.56	13.78	3.00	0.00

(b) The DTW distances of learning sequences over cases in $\mathbf{S}_l$

Table 3: The DTW distances of learning sequences over cases in each segment

In our context, leveraging the capabilities of DTW, we establish clusters that unveil latent patterns embedded within learning behaviors. Also, within the process of constructing clusters, diverse clustering techniques can be employed, including K-means

clustering (Krishna & Murty, 1999), Hierarchical clustering (Murtagh & Contreras, 2012), and DBSCAN (Schubert et al., 2017). Here, we apply K-means clustering, a widely used method with its simplicity and efficiency. Moreover, an essential step involves identifying the most suitable number of clusters, and a couple of methods can be considered, e.g., the silhouette score (Shahapure & Nicholas, 2020) and elbow method (Syakur et al., 2018). To achieve this, we employ the silhouette score (Shahapure & Nicholas, 2020), a widely used and valuable tool for this task, as measured with the following formula

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

, where $a(i)$ is the distance to other data points within the same cluster, and $b(i)$ is the smallest average distance to data points in other clusters. As explained in the formula, the silhouette score assesses the quality of clusters by measuring the cohesion and separation between them, enabling us to specify the optimal number of clusters that best captures the inherent patterns in the data.

After applying clustering to a dataset comprising time series, we identify a barycenter, referred to as a centroid, for each cluster (Petitjean et al., 2011). These barycenters effectively stand as the central or prototype time series within their clusters, capturing the common behavior of the time series in that group. Note that it is crucial to compute these barycenters while considering the DTW mechanism instead of a simple arithmetic mean to seize the temporal nuances in the time series data accurately. Let μ is the barycenter of a set of the sequences D using the DTW mechanism. The formula

$$\min_{\mu} \sum_{x \in D} DTW(\mu, x)^2 \quad (3)$$

is the objective of the optimization problem to derive the barycenter. To this end, the DTW Barycenter Averaging algorithm (Petitjean et al., 2011) readily facilitates the identification of barycenters. These barycenters, each corresponding to a cluster, play a pivotal role in the subsequent step of mapping the clusters to pre-defined learning cohorts.

In our running example, we can also apply clustering based on the DTW distances in Table 3. Figure 5 illustrates the clustering outcomes for two segments, S_s and S_l .

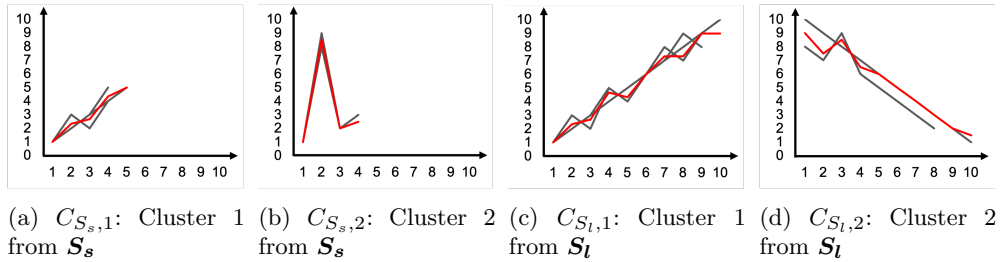


Figure 5: An example of clustering results from S_s and S_l

As a result, we identify two clusters for each segment. In segment $\mathbf{S}_s$, we observe the mixture of cases c2 and c4 into a singular cluster ($C_{S_s,1}$), while the remaining cluster ($C_{S_s,2}$) is composed of c6, c8, and c10. Likewise, in segment $\mathbf{S}_l$, three cases, c1, c3, and c5, form a single cluster ($C_{S_l,1}$), while c7 and c9 comprise the other cluster ($C_{S_l,2}$). Additionally, when considering the data within each cluster, we can compute a sequence of centroids to represent the data. In Figure 5, the red line represents the barycenters of each cluster. Details are as follows: $\mathbf{b}_{C_{S_s,1}} = (1.0, 8.5, 2.0, 2.5)$, $\mathbf{b}_{C_{S_s,2}} = (1.0, 2.3, 2.7, 4.3, 5.0)$, $\mathbf{b}_{C_{S_l,1}} = (1.0, 2.3, 2.7, 4.7, 4.3, 6.0, 7.3, 7.3, 9.0, 9.0)$, and $\mathbf{b}_{C_{S_l,2}} = (9.0, 7.5, 8.5, 6.5, 6.0, 5.0, 4.0, 3.0, 2.0, 1.0)$. It is evident that these barycenters effectively capture the characteristics of the time-series data within each cluster.

4.3 Mapping clusters to learning cohorts

This section explains mapping the clusters obtained through our time-series-based clustering approach to the predefined learning cohorts. As discussed, our cohorts are defined by four dimensions, e.g., engageness, direction, repetitiveness, and orderliness. Therefore, here, we present a comprehensive understanding of how we establish criteria and employ methods to subdivide and differentiate learning patterns within each dimension. Note that the direction and repetitiveness are analyzed by a unified approach; thus, we provide three distinct methods among four. Also, we utilize the barycenters of clusters as an input to map them to the cohorts.

4.3.1 Evaluation of the engageness

Foremost, we use the lengths of barycenters to evaluate the engageness of each cluster. In this regard, we first analyze the distribution of learning record lengths, gaining insights into the diverse engagement patterns among learners. To describe the distribution, it is available in measuring centers, e.g., mean or median, and spreads, e.g., the first and third quantiles. Next, we determine an appropriate threshold for categorizing learners as engagers or observers based on the distribution of the learning lengths. In particular, we should notice the issue of brief records, characterized by a tiny percentage of engaging in learning materials, such as less than 10% of the total course content. These brief records challenge the accurate identification of learning patterns and their categorization into cohorts. That is, they need to be excluded from the analysis.

In our running example, the length of barycenters for both clusters within the short segment ($\mathbf{b}_{C_{S_s,1}}, \mathbf{b}_{C_{S_s,2}}$) is 5, whereas in the longer segment ($\mathbf{b}_{C_{S_l,1}}, \mathbf{b}_{C_{S_l,2}}$), this length is 10. Hence, we can determine that the clusters with shorter barycenters represent observer behavior, while the clusters with longer barycenters correspond to engager behavior.

4.3.2 Evaluation of the direction and repetitiveness

Regarding the direction and repetitiveness, we introduce the Mann-Kendall test (Kendall, 1948; Mann, 1945). The Mann-Kendall test is a non-parametric statistical test used to determine whether there is a monotonic trend in a time series dataset. That is, it is beneficial to identify the trend of slopes, either increasing or decreasing, and in our context, it represents whether learners follow the guidelines. To perform

the hypothesis testing, we should compute the Mann-Kendall test statistic Z_{MK} , derived by the Mann-Kendall statistic S and the variance $V(S)$. First, we compute the Mann-Kendall statistic S for a time series $\mathbf{x} = (x_1, x_2, \dots, x_n)$ with the following formula

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_i - x_j) \quad (4)$$

, where n is the length of the sequence $\mathbf{x}$ and sgn is the indicator function. Specifically, it calculates the difference between the values at two different time points, i.e., x_i, x_j , using the formula

$$\text{sgn}(x_i - x_j) = \begin{cases} 1 & x_i - x_j > 0 \\ 0 & x_i - x_j = 0 \\ -1 & x_i - x_j < 0 \end{cases} \quad (5)$$

, which shows they are positive, negative, or zero. Also, the variance $V(S)$ is computed with

$$V(S) = \frac{1}{18} \left(n(n-1)(2n+5) - \sum_{k=1}^p t_k(t_k-1)(2t_k+5) \right) \quad (6)$$

, where p is the number of tied groups and t_k is the number of observations contained in the k th tied group. Next, we compute the Mann-Kendall test statistic Z_{MK} based on the S and $V(S)$:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{V(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{V(S)}} & S < 0 \end{cases} \quad (7)$$

, which ensures that Z_{MK} follows an approximately normal distribution for sufficiently large samples sizes. Note that a positive value of Z_{MK} holds that the data tends to increase over time, while a negative value indicates a tendency for the data to decrease with time. Then, we compare it to critical values to test the null hypothesis H_0 , i.e., there is no monotonic trend, versus the alternative hypothesis H_a , i.e., there exists either an upward or a downward monotonic trend. At the type I error rate α of the test, H_0 is rejected and H_a is accepted if $|Z_{MK}| \geq Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ denotes the $100(1 - \alpha/2)^{th}$ percentile of the standard normal distribution, and $|\cdot|$ is the absolute value function. Based on these results, we can categorize learners in forward and backward directions.

In addition to the positive or negative slope itself, we also identify the steepness of the slope. Basically, the positive slope indicates that the learner is progressing through learning materials as expected, while the negative implies that the learner does not follow the guidelines. The steepness of the slope is beneficial for analyzing the skipped, regulated, or repeated learning behaviors. In other words, when the slope is near 1.0, it signifies that learners adhere to the pre-determined sequence of learning materials. A slope greater than 1.0 suggests that the learners may skip some learning materials, while a slope below 1.0 implies that they revisit certain materials, indicating repetitiveness. To identify the steepness of the slope, we employ Theil-Sen estimator (Sen, 1968; Theil, 1950), a non-parametric method for estimating the slope by robustly fitting a line to data points. Let $\mathbf{x} = (x_1, x_2, \dots, x_i, x_j, \dots, x_n)$ be the time series

dataset at time points $t_1, t_2, \dots, t_n$. The Theil-Sen estimator Q is computed by the following formula

$$Q = \text{med} \left(\frac{x_j - x_i}{t_j - t_i} \right) \quad (8)$$

, where med is the median function to find the middle number in a set of values. Therefore, we can evaluate the steepness by calculating the median of the slopes determined by all pairs of data points.

Applying our running example to these methods, we can compute the Mann-Kendall test statistics as follows: 0.34, 2.20, 3.61, and -3.65 for each barycenter ($\mathbf{b}_{C_{S_s,1}}$, $\mathbf{b}_{C_{S_s,2}}$, $\mathbf{b}_{C_{S_l,1}}$, $\mathbf{b}_{C_{S_l,2}}$). Considering the critical values of 1.96 and -1.96 at a significance level of 0.05, we can conclude that two barycenters $\mathbf{b}_{C_{S_s,2}}$ and $\mathbf{b}_{C_{S_l,1}}$ exhibit a significant increasing trend, while $\mathbf{b}_{C_{S_l,2}}$ shows a significant decreasing trend. Additionally, the Theil-Sen estimator values are calculated as follows: 0.5, 1.0, 0.9, and -0.9 for each barycenter ($\mathbf{b}_{C_{S_s,1}}$, $\mathbf{b}_{C_{S_s,2}}$, $\mathbf{b}_{C_{S_l,1}}$, $\mathbf{b}_{C_{S_l,2}}$). Consequently, we can conclude that a single barycenter $\mathbf{b}_{C_{S_s,1}}$ represents recurrent behaviors, while the others do not.

4.3.3 Evaluation of orderliness

Lastly, two techniques are available for evaluating the level of orderliness: the log-difference approach and the Pettitt test (Pettitt, 1979). These two methods are capable of detecting sudden shifts or fluctuations in the data. The log-difference approach computes the squared values of logarithmic differences between consecutive data points, i.e., $(\log(\frac{x_i}{x_{i-1}}))^2$ for $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Thus, it quantifies the rate of change between data points. Then, if the squared values of log difference exceeds a certain threshold, it indicates a high degree of fluctuation or variability. The Pettitt test is a statistical method to detect significant shifts in data. To test the null hypothesis H_0 , i.e., a change point does not exist, against the alternative H_a , i.e., a change point exists in $\mathbf{x} = (x_1, x_2, \dots, x_i, x_j, \dots, x_n)$, we measure the Pettitt test statistic K_n as follows:

$$K_n = \max |U_{t,n}| \quad (9)$$

, where

$$U_{t,n} = \sum_{i=1}^t \sum_{j=t+1}^n \text{sgn}(x_i - x_j). \quad (10)$$

Here, sgn is the function defined in Equation 5. The approximate significance probability of K_n is computed with

$$p \simeq 2 \exp \left(\frac{-6K_n^2}{n^3 + n^2} \right). \quad (11)$$

If the p-value is lower than the significance level, it suggests that breakpoints exist in the time series $\mathbf{x}$. The value x_t at time t is considered the most significant change in the data.

Lastly, based on our running example, we computed the square values of log differences between consecutive data points in barycenters for each cluster ($\mathbf{b}_{C_{S_s,1}}$, $\mathbf{b}_{C_{S_s,2}}$,

$b_{C_{S_l,1}}$, and $b_{C_{S_l,2}}$). By considering a threshold of 1.0, we identified that there are two fluctuations in $b_{C_{S_s,1}}$, while the other do not involve any fluctuations.

4.3.4 Mapping based on the evaluation results

Lastly, on the basis of the evaluation results of barycenters, we proceed to classify learners into distinct learning cohorts within each cluster. To accomplish this, we employ a systematic classification approach, as detailed in Table 1

Table 4 offers a summary of the evaluation results for the four clusters in our running example, along with the associated learning cohorts for each cluster.

Table 4: A summary of mapping results in our running example

Cluster	Case	Evaluation				Cohort		
$C_{S_s,1}$	c2, c4	Observing	Forward	Recurrent	Disorderly	Recurrent	Disorderly	Observers
$C_{S_s,2}$	c6, c8, c10	Observing	Forward	-	Orderly	Orderly	Orderly	Observers
$C_{S_l,1}$	c1, c3, c5	Engaging	Forward	-	Orderly	Orderly	Orderly	Engagers
$C_{S_s,2}$	c7, c9	Engaging	Backward	-	Orderly	Orderly	Orderly	Contrary Learners

As illustrated in Table 4, it is evident that each cluster is assigned to a distinct learning cohort.

5 Case study

In this section, we present the application of our proposed method to datasets from a real-world online educational platform. Our evaluation’s core aim is to showcase the effectiveness and practicality of our method in identifying and categorizing learning cohorts within actual educational contexts.

To conduct an in-depth case study, we have established two research questions aimed at exploring both the method’s effectiveness in learner mapping and its capability to uncover insightful learning patterns:

- **RQ1:** How effectively does the proposed method categorize students into appropriate learning cohorts in a real-world context?
- **RQ2:** What insights into learner behaviors and characteristics can be derived from analyzing the identified learning cohorts?

Addressing **RQ1**, we apply our method to three diverse datasets, assessing its ability to discern and assign learners to their corresponding cohorts. This evaluation includes an analysis of learner distribution across cohorts and courses, as well as examining consistency in learning behaviors for students enrolled in multiple courses.

For **RQ2**, we analyze the specific behaviors and characteristics of each cohort by employing a series of performance indicators (Cho et al., 2017) and process mining techniques (Juhaňák et al., 2019; Van Der Aalst, 2012). These tools are adept at unveiling detailed learner interactions and learning patterns, offering a granular view of the unique characteristics that define each cohort. Through this analysis, we aim to

uncover the diverse learning dynamics present within the data, providing a foundation for targeted educational interventions.

The following subsections provide a detailed overview of the datasets used (Subsection 5.1) and articulate the findings in response to our research questions (Subsection 5.2 and Subsection 5.3).

5.1 Datasets

The dataset employed in this case study originates from an online education platform offering courses across a broad spectrum of academic disciplines, including software development, artificial intelligence, data science, programming language, and mathematics. Among the courses in these disciplines, we focus on three specific courses related to data science. The dataset holds a substantial timeframe, spanning approximately four years from March 2018 to May 2022. Also, to ensure consistency in our results, we only cover the records of video-watching activities. More detailed information regarding the dataset for each of the selected courses is provided in Table 5.

Table 5: The basic characteristics of the courses

Course	Course Information			Data Information		
	Discipline	#Chapters	#Lectures	#Students	#Events	#Events per student (Avg., Min., Max.)
A	Machine Learning	12	49	9,425	95,101	(10.09, 1, 97)
B	Deep Learning	4	47	4,062	32,844	(8.09, 1, 96)
C	Data Science	8	36	4,918	43,495	(8.84, 1, 55)

Course A, which focuses on machine learning, boasts a comprehensive curriculum with numerous chapters and lectures. It is the largest among the three courses in terms of both chapters and lectures. In contrast, Course B, centered around deep learning, comprises smaller chapters but a substantial number of lectures. Course C, dedicated to data science, has the fewest lectures among the three. Regarding learning data, Course A has the highest number of students, nearly 9,500, while Course B and Course C have 4,062 and 4,918 students, respectively. This trend is consistent with the volume of events. Course A includes over 95,000 events, while Course B and Course C have 32,844 and 43,495 events, respectively. Also, computing the average number of events per student, we find that Course A has an average of 10.09 events per student, Course B averages 8.09 events per student, and Course C has an average of 8.84 events per student. Note that we exclusively consider events related to video-watching activities completed according to the progress rate.

5.2 Learning cohorts for the three courses

First, we apply the proposed method to the datasets of three courses to derive the learning cohorts. To facilitate this process, we categorize the data into 6, 4, and 5 length groups for each course based on the descriptive statistics related to the lengths of the learning sequences. Detailed information regarding these length groups for each

course is presented in [Table 6](#). Note that the cases with brief records, i.e., those with a length of less than 5, are excluded and not included as subjects for mapping into learning cohorts. Next, we employ the K-means clustering with dynamic time warping for each length group within the three datasets. To determine the optimal number of clusters for each group, we utilize both silhouette scores and qualitative assessments. In particular, we visualize the clustering results with different candidate numbers of clusters, focusing on those with high silhouette scores. This visualization allows us to select the optimal number of clusters that capture distinct patterns within the data. [Table 6](#) presents the specifics, including the number of clusters for each group and the corresponding silhouette scores.

After deriving clusters for each length group, we compute the barycenter for each cluster, which is essential for mapping the clusters to the predefined learning cohorts. Following our approach, we consider four dimensions: engagement, direction, repetitiveness, and orderliness.

To assess the engagement perspective, we analyze the lengths of the barycenters, explicitly focusing on the maximum value within the range of each length group. To perform this assessment, we set a threshold of 50% of each course’s total number of lectures. If the length of the barycenter falls below this threshold, the learner is categorized as an observer. Conversely, the learner is categorized as an engager if the length surpasses the threshold. More in detail, in the case of courses A and B, any barycenter length up to 22 is designated as an observer, while for course C, any barycenter length up to 18 falls under the observer category.

Next, we employ the Mann-Kendall test with a significance level of 0.01 to assess the direction of learning behaviors. In the context of the Mann-Kendall test outcomes, a diminishing trend observed in the barycenter leads to the classification of the learner’s behavior as backward direction. Conversely, the indication of an ascending trend drives the classification of the learner as displaying a forward direction.

Regarding the repetitiveness and partial orderliness, we leverage the slope of the barycenter. When the slope aligns closely with 1.0, within the range of 0.7 to 1.3, it implies non-recurrent learning, where the learner follows the predetermined sequence. A slope below 0.7 indicates recurrent learning, implying that the learner repeatedly engages with the same materials. Conversely, a slope greater than 1.3 indicates disorderly learning, suggesting that the learner tends to skip or engage selectively with certain course materials.

Lastly, in our evaluation of orderliness, we leverage the log-difference approach, calculating the squares of log-differences between consecutive data points. Here, one of the values surpassing the threshold of 0.6 conveys a disorderly orderliness, denoting a significant degree of fluctuation in the learner’s interactions. Conversely, all values below 0.6 denotes an orderly state of orderliness, indicating a reduced fluctuation in the learner’s interactions with the learning materials.

Based on the evaluation results for the four dimensions of each cluster, we proceed to assign these clusters into cohorts, as illustrated in [Table 1](#). [Figure 6](#) provides a visual representation of each cohort’s learning behaviors and barycenters across the three courses. In each figure, the black lines represent the learning behaviors of the whole learners within the cluster, while the red line represents the barycenter, which

Table 6: The details of length groups and clustering results for each course

Course A			Course B			Course C		
range of length	# of clusters	silhouette score	range of length	# of clusters	silhouette score	range of length	# of clusters	silhouette score
5-7	8	0.69	5-9	7	0.61	5-8	5	0.65
8-9	7	0.69	10-22	6	0.43	9-18	7	0.41
10-13	14	0.46	23-47	6	0.40	19-27	5	0.42
14-22	13	0.38	48-96	5	0.39	28-36	6	0.60
23-49	6	0.32				37-55	4	0.62
50-97	3	0.41						

encapsulates the learning behaviors of the cohort. Note that the scales of the x-axis and y-axis are consistent across all figures except for [Figure 6\(k\)](#) and (i).

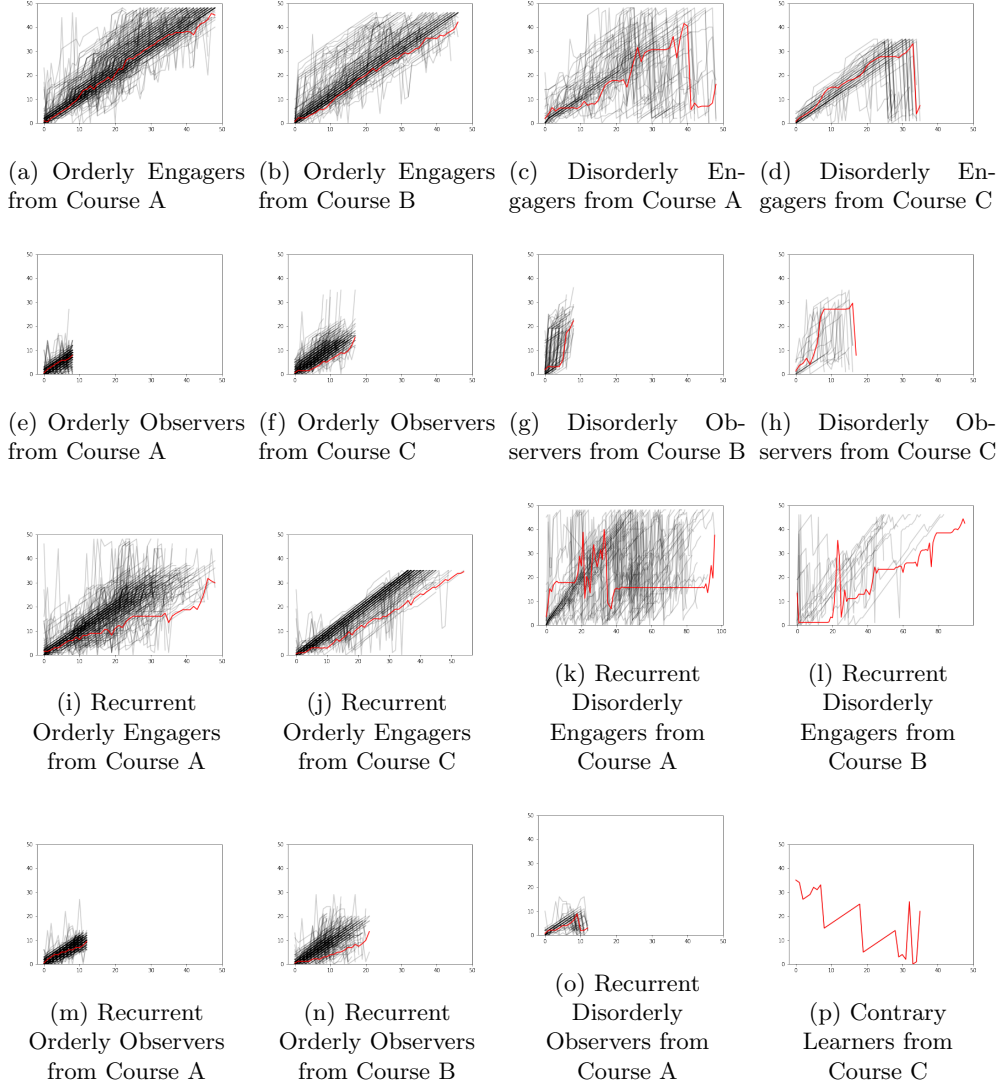


Figure 6: Learning behaviors and barycenters of each cohort

As described in [Figure 6\(a\)](#) and (b), orderly engagers predominantly adhere to the prescribed course sequence, completing it methodically. Their learning behavior is characterized by a slope close to 1.0, indicating conformity to the reference sequence. Disorderly engagers, described in [Figure 6\(c\)](#) and (d), are similar to orderly engagers in

terms of the number of lectures undertaken, while this cohort exhibits fluctuations in their learning approach. Despite some disarray, they manage to complete the majority of course materials. As depicted in Figure 6(e), (f), (g), and (h), orderly observers and disorderly observers have similar characteristics with orderly engagers and disorderly engagers; however, they discontinue their participation early and do not complete the entire course. In Figure 6(i) and (j), there are typical behaviors of recurrent orderly engagers, not only adhering strictly to the given guidelines but also revisiting course materials for review and understanding. In contrast, recurrent disorderly engagers, visualized in Figure 6(k) and (l), deviate from the reference sequence and engage in exploratory behaviors. They also revisit some course materials, resulting in a moderately sloped pattern. In Figure 6(m), (n), and (o), we can identify the behaviors of recurrent orderly observers and recurrent disorderly observers, which are similar to recurrently orderly engagers and observers; however, they discontinue the course early. Lastly, contrary learners, the rarely appeared cohort, have the shape of negative slope as depicted in Figure 6(p). That is, they take courses contrary to the given guidelines.

Following the mapping of clusters to learning cohorts, we clarify the distribution of learners among these cohorts across the three courses. Table 7 details the frequency of learners engaged in each cohort for three courses.

Table 7: Distribution of learners for each cohort

Cohort	Course A		Course B		Course C	
Orderly Engagers	209	2.2%	146	3.6%	408	8.3%
Disorderly Engagers	51	0.5%	-	-	80	1.6%
Orderly Observers	1,027	10.9%	107	2.6%	305	6.2%
Disorderly Observers	415	4.4%	130	3.2%	131	2.7%
Recurrent Orderly Engagers	423	4.5%	118	2.9%	206	4.2%
Recurrent Disorderly Engagers	482	5.1%	181	4.5%	129	2.6%
Recurrent Orderly Observers	1,563	16.6%	696	17.1%	730	14.8%
Recurrent Disorderly Observers	387	4.1%	131	3.2%	52	1.1%
Contrary Learners	-	-	-	-	1	0.0%
Under 10%	4,868	51.6%	2,553	62.9%	2,876	58.5%

The most prevalent cohort across all courses is the recurrent orderly observers, indicating that a substantial proportion of learners generally adhere to predetermined guidelines with recurring course materials but tend to discontinue their participation prematurely. Following this cohort, course A has a higher proportion of orderly observers, with over 10%. Conversely, course B has the second-largest proportion of recurrent disorderly observers; however, it comprises only 4.5% of the students. In the case of course C, orderly engagers also occupy a significant proportion, which is not prominently featured among the frequent cohorts in the other courses. Also, course C holds a single learner classified as the cohort of contrary learners. The least prevalent cohort is the disorderly engagers and recurrent disorderly observers across the three courses. Note that over half of learners engage with the course materials for less than 10% of the course materials.

We also investigate whether learners who participate in multiple courses show consistent learning behaviors across these courses. Our study employs the chi-squared

goodness of fit test. This statistical hypothesis test helps us assess the degree of difference between observed and expected values. The null hypothesis H_0 suggests no significant difference exists between the number of students holding the same learning behaviors across courses and those who do not, while the alternative hypothesis H_a proposes that these two values are different. To test H_0 against H_a , we compute the chi-squared test statistic χ^2 as follows:

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (12)$$

, where k is the number of categories, O_i is the i^{th} observed value, and E_i is the i^{th} expected value. At the significance level of α , H_0 is rejected and H_a is accepted if $\chi_{obs}^2 > \chi_{k-1;\alpha}^2$, where $\chi_{k-1;\alpha}^2$ denotes chi-square critical value calculated from chi-square distributions.

In our dataset, we have 670 students who are categorized into cohorts across two or more of the three courses. Among these students, 162 learners maintain the same cohort in at least two courses, while the remaining 508 have different cohorts across these courses. Based on these observed values and the expected values for each scenario of 335, we compute the chi-squared statistic as 178.68 based on the formula. Comparing this value with the critical value of 6.63 at a significance level of 0.01, we find that the chi-squared value significantly exceeds the critical value; thus, the p-value is less than 0.01. As a result, we reject the null hypothesis and accept the alternative hypothesis. This indicates a significant difference between the number of students with consistent learning behaviors across courses and those without. Specifically, the number of students with varying cohorts greatly exceeds those with consistent cohorts. Therefore, learning cohorts depend more on the specific courses rather than the individual learners themselves. Nevertheless, it is worth noting that 26% of the students maintain consistent cohorts in at least two courses.

5.3 Unveiled insights for learning cohorts

To delve deeper into the characteristics of learning cohorts, we conduct an extensive analysis. Initially, we establish specific performance indicators to assess learners' academic achievements and subsequently compare these indicators across cohorts. Our evaluation encloses two categories of indicators: quiz and project scores, and the pass ratio. Quiz and project scores denote the students' performance in quizzes or project assignments, quantified as a proportion of total achievable scores. Courses A and C only hold a series of quizzes, while course B entails a couple of individual projects focused on deep learning. Also, each course features specific criteria for assessing whether a student successfully passes or fails the course. For instance, courses A and C consider quiz scores and the progress made in watching video lectures, whereas course B relies on project scores and the progression of video lectures. Based on these criteria, we compute the pass ratio to measure students' learning achievements within each cohort. [Table 8](#) describes the details of learning achievements across learning cohorts.

Table 8: Quiz and project scores and pass ratio for each cohort

Cohort	Quiz and Project Score			Pass Ratio		
	Course A	Course B	Course C	Course A	Course B	Course C
Orderly Engagers	0.165	0.186	0.459	0.167	0.151	0.468
Disorderly Engagers	0.039	-	0.475	0.039	-	0.488
Orderly Observers	0.007	0.034	0.056	0.007	0.028	0.056
Disorderly Observers	0.010	0.032	0.231	0.010	0.023	0.244
Recurrent Orderly Engagers	0.232	0.027	0.625	0.234	0.017	0.636
Recurrent Disorderly Engagers	0.109	0.192	0.535	0.110	0.149	0.550
Recurrent Orderly Observers	0.002	0.012	0.022	0.002	0.010	0.022
Recurrent Disorderly Observers	0.016	0.032	0.096	0.016	0.023	0.096
Contrary Learners	-	-	1.000	-	-	1.000
Total	0.046	0.057	0.241	0.046	0.044	0.246

Considering the inherent characteristics of online education platforms, it is evident that the overall learning achievements of learners tend to be relatively low. Also, as previously mentioned, determining course pass or fail is closely tied to quiz and project scores. Consequently, these two metrics exhibit similar trends. However, variations become apparent when examining these metrics across cohorts within each course and across different courses. Regarding both assignment scores and pass ratios, we observe that recurrent orderly engagers achieve the highest scores in courses A and C. In contrast, course B shows the highest values for recurrent disorderly engagers. This divergence can be attributed to the design of the courses. In courses A and C, where quizzes are closely integrated with video lectures, students who diligently follow the video lectures with adherence to the guidelines tend to attain better academic achievements. However, in course B, projects are spread across the lectures, and adherence to a specific order does not hold the same significance, resulting in different cohorts achieving higher performance.

Furthermore, these learning management systems also capture learner course interaction data. These interactions encompass both active and passive participation. Active participation involves posting on discussion boards or leaving comments, whereas passive participation includes clicking the recommend button. To quantify these interactions, we establish an interaction metric. This metric is computed by summing active and passive participation scores, with a relatively low weight assigned to passive participation. Specifically, we assign a weight of 0.2 to passive participation, considering the differing volumes of these two types of engagement. [Table 9](#) introduces the results of interaction scores for each cohort.

On average, the interaction scores across all courses fall within the range of 0.069 to 0.278, indicating a generally low degree of interaction. At the cohort level, students classified as recurrent orderly engagers have the highest interaction scores in all three courses. This observation suggests that students with higher engagement levels tend to have more interactions with their peers in these courses. Additionally, when considering both [Table 8](#) and [Table 9](#), it becomes evident that cohorts with higher interaction scores, such as recurrent disorderly engagers in Courses A and C, also tend to achieve higher learning outcomes. Hence, it can be inferred that there is a positive correlation between interaction scores and learning achievements.

Table 9: Interaction scores for each cohort

Cohort	Course A	Course B	Course C
Orderly Engagers	0.101	0.371	0.393
Disorderly Engagers	0.075	-	0.380
Orderly Observers	0.041	0.133	0.241
Disorderly Observers	0.039	0.108	0.137
Recurrent Orderly Engagers	0.202	0.608	0.578
Recurrent Disorderly Engagers	0.147	0.531	0.487
Recurrent Orderly Observers	0.039	0.172	0.138
Recurrent Disorderly Observers	0.033	0.089	0.069
Contrary Learners	-	-	0.000
Total	0.069	0.253	0.278

Next, we analyze process models of the cohorts to understand the difference in their learning processes. This analysis offers a comprehensive view, allowing us to discern and define the distinguishing features of each cohort at a glance. To accomplish this, we collect event logs from four specific cohorts within course A, specifically orderly engagers, recurrent orderly engagers, orderly observers, and disorderly engagers, as depicted in Figure 6(a), (c), (e), and (i). We then leverage the Fluxicon DISCO toolkit (Günther & Rozinat, 2012) to generate process models based on these logs. Figure 7 present the discovered process models for partial cohorts from Course A.

In this context, we apply consistent thresholds for the quantity of nodes and edges to ensure an accurate comparison among cohorts. Specifically, we enclose all nodes, representing 100% of activities, while maintaining a relatively lower number of edges, equating to 20% of paths, to facilitate a more transparent comprehension of process models. As illustrated in Figure 7(a) and (c), it becomes evident that orderly engagers and observers hold relatively straightforward process models, with few exceptions, as they diligently adhere to the predefined guidelines. However, the length of these process models varies following the level of engagement by each cohort. Also, comparing Figure 7(a) to Figure 7(b), the latter appears more complicated due to its recurrent investigations involving a return to previous lectures. Furthermore, it is discernible that Figure 7(d) stands out as the most intricate process model among the four. This complexity is attributed to the self-regulated behaviors of disorderly engagers, resulting in diverse flows and back-and-forth interactions in their learning journey.

In this manner, we can assess the identified learning cohorts by investigating the corresponding process models. These models, as they are constructed from data, are valuable for comprehending learner behaviors. However, when the model is constructed using data from the whole learners, it tends to become exceedingly complex, which can obscure the detection of significant hidden or inherent patterns or behaviors. Put differently, when the process discovery is employed in conjunction with the learning cohort derivation method introduced earlier, it can synergize efforts to uncover various learner behavior patterns.

6 Discussion

This section includes an in-depth discussion of our approach and the outcomes of the case study. Initially, we present the findings related to the research questions in

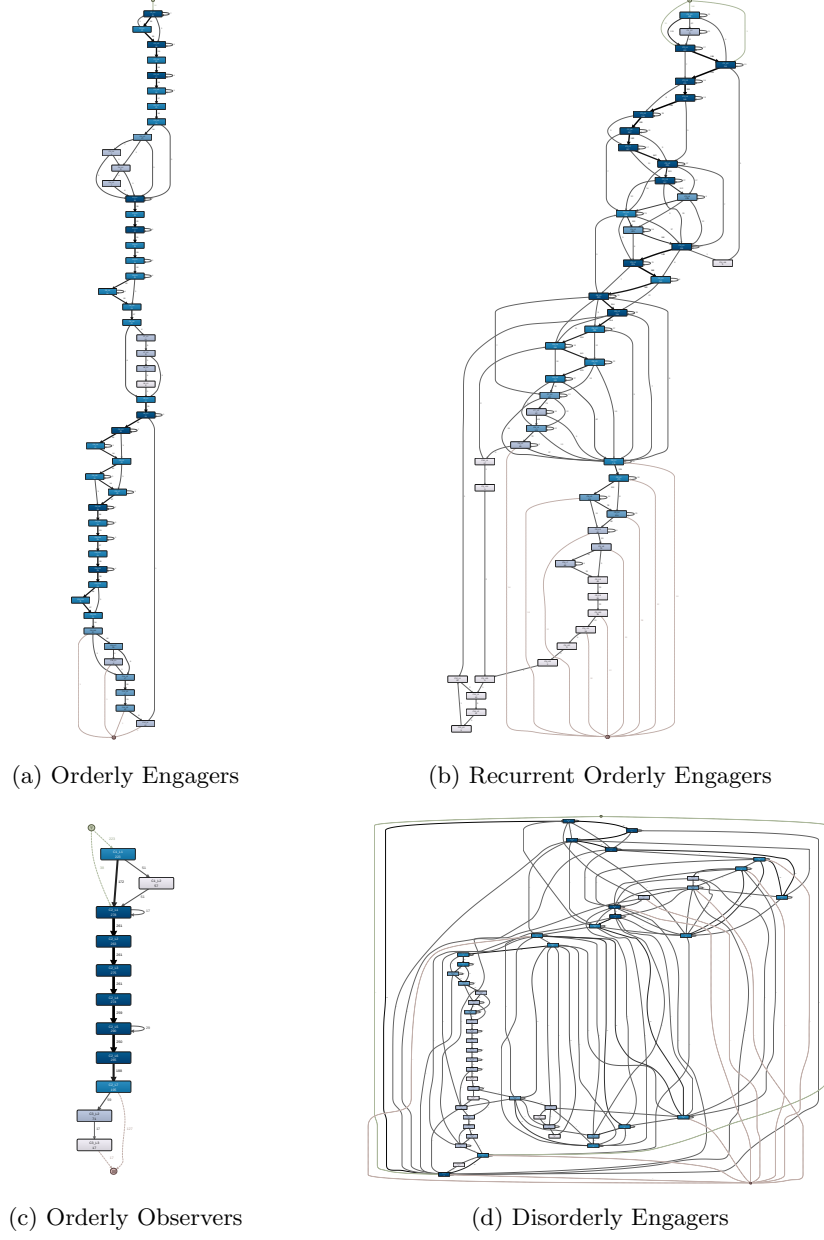


Figure 7: Process models for partial cohorts from Course A

[Subsection 6.1](#). Subsequently, in [Subsection 6.2](#), we present treats to the validity of our findings. Next, we provide practical applications of our proposed approach in

[Subsection 6.3](#). Finally, [Subsection 6.4](#) describes the limitations of our approach and offers directions for future research.

6.1 Results concerning the research questions

The first research question revolves around validating and evaluating our approach for categorizing learners into distinct learning cohorts. Our findings demonstrate that our approach effectively derives these learning cohorts using the available datasets. Going into greater detail regarding the results, we observe variations in the distribution of learners across learning cohorts for each course. This non-consistency in learning cohort distribution can be attributed to the unique characteristics of each course.

In the cohort context, we note that, on the whole, there is a higher proportion of recurrent learners compared to non-recurrent ones. Additionally, there are more orderly learners than disorderly ones, while in comparison to engagers, there are more observers among the learners. This indicates that learners in online platform systems commonly exhibit specific characteristics. These tendencies include repeatedly engaging with course materials, adhering to the provided guidelines, and exploring a relatively limited portion of the course materials. This observation aligns with findings from previous research (Andergassen et al., 2014; Haar & Bell, 2020; Monaghan & Rowson, 2008), which has highlighted the significance and efficacy of iterative learning in both online and offline environments. In addition to this, while there is a demand for self-regulated learning based on planning, problem-solving, and self-evaluation in online settings (Zimmerman, 2002), the effectiveness of regular study behaviors by keeping pace with the class schedule is also being discussed (Asarta & Schmidt, 2013; You, 2016). Also, it is worth noting that a significant portion of learners, exceeding 50% for each course, falls into the category of those who discontinue their engagement and abandon the courses after completing only a few lectures. This underscores the high attrition rates in online learning platforms, where many learners do not progress beyond the initial stages of a course. Regarding this, early dropout in online platforms, including MOOCs, has become a significant issue, and numerous studies have been undertaken to identify the factors contributing to dropout (Goopio & Cheung, 2021; Lu et al., 2017) and to predict its occurrence (Charitaki et al., 2023).

Based on the statistical results of exploring the learning cohorts across courses, it becomes apparent that learners have different behaviors depending on the specific course they are enrolled in. This implies that the alignment between learning cohorts and courses is more substantial than that with individual learners. In light of these findings, when striving to enhance learning behaviors, it becomes imperative to consider course-specific modifications to the structures and materials. However, it is of importance to note that a significant proportion of learners, approximately 26% in our dataset, possess their own distinct characteristics in terms of learning behaviors. This suggests the need for a learning support system that can provide personalized support tailored to the individual learning behaviors of these students. In relation to this, research has also been focused on the effectiveness, role, and necessity of student-centered learning platforms (Kurhila et al., 2004; Serban & Vescan, 2019). Furthermore, there is ongoing research into methods for measuring the student-centered qualities and user

experience of the learning management systems (Malm & Defranco, 2012; Santoso et al., 2016).

Regarding the second research question, we conduct a detailed exploration of learning cohorts to understand their characteristics, focusing on their relationship with learning achievements and interactions. Our findings reveal some critical insights. First, across the courses, we observe differences in overall scores, which can be attributed to variations in the difficulty levels of the courses. However, a general trend emerges - recurrent orderly engagers tend to have the highest learning achievements. This suggests that consistently engaging with course materials and following guidelines are crucial in determining learners' success. Furthermore, engagers, in general, achieve higher scores compared to observers. This highlights the significance of active engagement in the learning process for achieving better outcomes. Consistent with our findings, numerous studies have documented a positive correlation between repetition and improved learning outcomes (Andergassen et al., 2014; Ledermüller & Fallmann, 2017). Also, research efforts have focused on developing various supports and strategies aimed at enhancing the learning outcomes of students with relatively low self-regulated learning abilities (Long & Alevan, 2017; Van den Boom et al., 2007; van Alten et al., 2020). Also, we investigate the relationship between learning cohorts and their interaction levels. The results indicate notable differences in interaction scores among learning cohorts within each course. While some studies have analyzed learners' patterns based on their interactions (Çebi & Güyer, 2020), there has been no research that examines the relationship between learner interaction and learning patterns in cohorts as this study does. Nonetheless, given the existing research that demonstrates the relationship between interaction and learning outcomes (Kang & Im, 2013; Lin et al., 2017), this finding opens up possibilities for tailoring strategies based on a student's specific learning cohort to improve learning achievements.

Furthermore, we generate process models to gain a more detailed perspective on the sequences of activities within each learning cohort. Specifically, we observe that orderly learners exhibit more straightforward sequences in their process models, whereas disorderly learners have complex, spaghetti-like process models. Additionally, the length of process models varies based on whether the cohort consists of engagers or observers. Moreover, recurrent learning characteristics tend to make process models more intricate, with back-and-forth flows indicating a higher level of revisiting and engagement. These process models offer valuable insights into different cohorts' behaviors and learning patterns. To this end, this study offers value by introducing a new approach in the process mining domain, where we construct learning cohorts based on educational contexts and derive process models for them, instead of applying traditional trace clustering (Song et al., 2009). Furthermore, other process mining techniques potentially uncover even more nuanced insights into the behavioral patterns of these cohorts, further enriching our understanding of learner behaviors.

In addition to these approaches, we can explore further analyses to gain a more comprehensive understanding of learning cohorts. One avenue is to leverage the features of learners, such as demographics and past academic performance, to investigate the characteristics of these cohorts. This approach allows us to establish connections between learners' attributes and their corresponding learning cohorts. Expanding on

the results of this method, it is possible to construct predictive models for learning cohorts based on these learner features. Such a model could serve as a valuable tool for providing personalized learning guidelines to newcomers, tailored according to the predicted learning cohorts. This way, we can enhance the overall learning experience and outcomes for diverse learners.

6.2 Threats of validity

Our case study, while providing insightful findings, also presents certain threats to validity that warrant careful consideration.

6.2.1 Internal threats

First, our study’s datasets are drawn from an online educational platform, focusing specifically on courses related to data science. The selection of these courses was guided by their relevance to our research questions and the availability of comprehensive data. However, this selection process may inherently introduce bias, as it might not represent the full spectrum of online learning behaviors across all possible domains.

Second, a significant challenge in evaluating our proposed method is the absence of a ground-truth for learning cohorts within the dataset. This lack of an established benchmark complicates the task of conclusively validating the accuracy and effectiveness of our cohort identification and categorization process. In our study, we rely on a qualitative validation of the coherence and relevance of the identified cohorts.

Third, the integrity of our findings heavily relies on the quality and completeness of the dataset. While we have endeavored to utilize a dataset that is both robust and representative, there is always the potential for missing data points or inaccuracies within the dataset itself. Such issues could potentially influence the identification of learning cohorts and the insights derived from them.

6.2.2 External threats

The external validity of our study is impacted by the cultural and demographic characteristics of the learners in the dataset. Online learning cohorts can exhibit behaviors influenced by their cultural backgrounds, educational systems, and personal circumstances.

Next, the learning behaviors observed and the resulting cohorts identified are influenced by the design, tools, and interaction opportunities provided by the specific online platform from which the data was collected. Different platforms may facilitate different types of engagement and learning behaviors, thus affecting the generalizability of our results.

Moreover, our study is based on data spanning four years, a period that has seen significant evolutions in online learning methodologies, tools, and student engagement strategies (Dhawan, 2020). While this timeframe provides a rich dataset, the rapidly evolving nature of online education may affect the long-term generalizability of our findings. Continuous research efforts are needed to adapt and validate the proposed method against newer datasets.

6.3 Applications

The practical applications and impact of our proposed approach within the online education sphere are extensive and address several pivotal challenges. This approach’s ability to categorize learners into well-defined cohorts based on their learning behaviors is a cornerstone for enhancing personalized learning experiences. This capability is critical in online education, where learners’ preferences and learning styles can vary widely (Bernacki et al., 2021). Personalized learning paths crafted from the insights provided by our approach can significantly improve learning outcomes and student satisfaction, as tailored content and instructional strategies cater directly to the needs of each cohort.

Another significant application of our approach lies in its potential to improve retention rates. By enabling the early identification of learners at risk of disengagement or dropout, educators and institutions can intervene proactively (Lee & Choi, 2011). Targeted support and resources can be allocated to those in need, potentially keeping more students engaged with their courses until completion. This aspect of the approach could be instrumental in addressing the high dropout rates often seen in online learning environments.

Furthermore, our approach offers valuable insights that can inform course design and improvement. Understanding how different cohorts interact with course materials allows for the optimization of content, structure, and delivery methods. This adaptive approach to course design ensures that educational offerings are more closely aligned with the varied needs of the learner population, potentially leading to better engagement and learning outcomes.

Lastly, the construction of predictive models based on the approach opens up possibilities for anticipatory support for newcomers. By leveraging learner features, educators can predict which cohort a new learner is likely to belong to. This prediction enables the provision of customized learning guidelines right from the start of a learner’s journey, offering a personalized roadmap that enhances their chances of success.

6.4 Limitations

This work also has several limitations. First, while our approach is highly automated, it requires some level of human involvement for specific decision-making processes. For instance, determining the appropriate number of clusters using the silhouette score and setting the threshold for the sum of squares of log differences at 0.6 still rely on human judgment and input. These decisions, crucial to the accurate identification and categorization of learning cohorts, depend significantly on expert judgment and interpretation of the data.

Furthermore, the computational complexity associated with the clustering process, particularly when employing techniques like dynamic time warping, poses challenges. Dynamic time warping, while effective in capturing the nuanced similarities between time-series data, is computationally intensive, especially with large datasets. This complexity can limit the method’s scalability and efficiency.

Another critical limitation is the static nature of cohort definitions over time. The approach, as designed, does not account for the evolution of learning behaviors over the duration of a course or program. Learner engagement and interaction with course materials may change, necessitating a dynamic approach to cohort identification that can adapt to these changes over time.

Lastly, there remains room for learning behaviors to be subdivided based on other aspects. For example, considering the actual time of learning events rather than the logical time can provide more detailed insights into how learners engage with course materials over time. This can help distinguish between those who engage consistently over a longer duration and those who may exhibit intense but shorter bursts of activity.

7 Conclusion

In this work, we introduced a novel approach for categorizing learners into meaningful cohorts based on diverse learning behaviors within the context of online education. By integrating four critical dimensions, i.e., engagement, direction, repetitiveness, and orderliness, our approach offers a comprehensive method to automatically map learners to distinct cohorts. This approach leverages a time-series-based clustering method, utilizing dynamic time warping and statistical techniques like the Mann-Kendall test and Theil-Sen estimator, to identify and categorize learning patterns effectively.

The application of our approach to datasets from three data science courses highlights its utility in providing detailed insights into learner behaviors and facilitating the development of personalized learning paths. Through the identification of distinct learning cohorts, we can better understand learners' interactions with course materials, paving the way for interventions tailored to specific learning needs.

Our work has significant implications from both academic and practical standpoints. First, our research represents a pioneering effort in constructing learning cohorts by considering a range of behavioral factors related to learning. These aspects hold the potential for further exploration in future research endeavors focused on understanding and analyzing student learning behaviors. Additionally, the learning cohorts we have developed can be valuable for categorizing learners in scenarios where cohort-based investigations are necessary, whether in qualitative or quantitative research settings. This opens up new roads for studying learning patterns and behaviors in educational contexts.

Furthermore, we introduce a novel perspective by conceptualizing learning behaviors as time-series data within a logical time frame. This representation allows for a more granular understanding of learning behaviors, including factors such as orderliness, engagement levels, and more. Also, leveraging this data, we apply time-series-based clustering techniques to enhance the clarity of learning cohorts and their subsequent mapping into learning clusters. This novel approach provides a fresh lens for a more comprehensive exploration of learning behaviors from various angles and perspectives, enriching the field of educational learning analytics.

Finally, our approach is characterized by its systematic and transparent nature, making it readily applicable in practical settings for identifying learning cohorts from specific datasets. Moreover, within a learning management system, our approach holds

the potential to serve as a foundational component for personalized support mechanisms. This versatility allows for its integration into various learning support systems, fostering the possibility of numerous case studies and developing best practices within the educational industry.

There are multiple directions for future work. First, to address the dependency on human input and enhance the framework's autonomy, future work could explore more automated, data-driven approaches. Leveraging machine learning algorithms could help automate the decision-making processes involved in cluster analysis and threshold settings. Second, to mitigate the computational demands of dynamic time warping, investigating more computationally efficient algorithms or approximations is crucial. Such advancements would help maintain the method's accuracy while enhancing scalability for analyzing extensive educational data. Third, incorporating temporal dynamics into cohort identification is another critical area for future work. A dynamic approach would allow the framework to adapt to changes in learner behaviors over time, providing a more accurate and reflective analysis of learner engagements. Fourth, extending our approach to cover temporal engagement patterns by considering actual time metrics of learning events offers another promising direction. This could enable a more detailed understanding of how learners engage with course materials over time.

Data Availability

Raw data supporting this study are not publicly available to preserve individuals' privacy.

Ethics declarations

Competing interests. The authors have no conflicts of interest to declare that are relevant to the content of this article and journal.

References

- Abdallah Moubayed, A. S., Mohammadnoor Injadat, & Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. *American Journal of Distance Education*, 34(2), 137–156. <https://doi.org/10.1080/08923647.2020.1696140>
- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering—a decade review. *Information systems*, 53, 16–38.
- Andergassen, M., Mödritscher, F., & Neumann, G. (2014). Practice and repetition during exam preparation in blended learning courses: Correlations with learning results. *Journal of Learning Analytics*, 1(1), 48–74.
- Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, 60, 383–398.
- Asarta, C. J., & Schmidt, J. R. (2013). Access patterns of online materials in a blended course. *Decision Sciences Journal of Innovative Education*, 11(1), 107–123.

- Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)? *Educational Psychology Review*, 33(4), 1675–1715.
- Bharara, S., Sabitha, A. S., & Bansal, A. (2018). Application of learning analytics using clustering data mining for students' disposition analysis. *Educ. Inf. Technol.*, 23(2), 957–984. <https://doi.org/10.1007/S10639-017-9645-7>
- Cao, Y., Gao, J., Lian, D., Rong, Z., Shi, J., Wang, Q., Wu, Y., Yao, H., & Zhou, T. (2018). Orderliness predicts academic performance: Behavioural analysis on campus lifestyle. *Journal of The Royal Society Interface*, 15(146), 20180210.
- Caspari-Sadeghi, S. (2022). Applying learning analytics in online environments: Measuring learners' engagement unobtrusively. *Frontiers in Education*, 7, 840947.
- Çebi, A., & Güyer, T. (2020). Students' interaction patterns in different online learning activities and their relationship with motivation, self-regulated learning strategy and learning performance. *Education and Information Technologies*, 25(5), 3975–3993.
- Charitaki, G., Andreou, G., Alevriadou, A., & Soulis, S.-G. (2023). A nonlinear state space model predicting dropout: The case of special education students in the hellenic open university. *Education and Information Technologies*, 1–18.
- Cho, M., Song, M., Comuzzi, M., & Yoo, S. (2017). Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques. *Decision Support Systems*, 104, 92–103.
- Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, 18(6), 683–695.
- Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.
- Danielsson, P.-E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, 14(3), 227–248.
- Deng, R., Benckendorff, P., & Gannaway, D. (2020). Linking learner factors, teaching context, and engagement patterns with MOOC learning outcomes. *J. Comput. Assist. Learn.*, 36(5), 688–708. <https://doi.org/10.1111/JCAL.12437>
- Dhawan, S. (2020). Online learning: A panacea in the time of covid-19 crisis. *Journal of educational technology systems*, 49(1), 5–22.
- Dong, L., & Kang, Y. (2022). Cultural differences in mindset beliefs regarding mathematics learning. *Current Opinion in Behavioral Sciences*, 46, 101159. <https://doi.org/https://doi.org/10.1016/j.cobeha.2022.101159>
- El-Sabagh, H. A. (2021). Adaptive e-learning environment based on learning styles and its impact on development students' engagement. *International Journal of Educational Technology in Higher Education*, 18(1), 53.
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6), 304–317.
- Ferguson, R., & Clow, D. (2015). Examining engagement: Analysing learner sub-populations in massive open online courses (moocs). *Proceedings of the fifth international conference on learning analytics and knowledge*, 51–58.

- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1), 59–109.
- García-Peñalvo, F. J. (2020). Learning analytics as a breakthrough in educational improvement. *Radical Solutions and Learning Analytics: Personalised Learning and Teaching Through Big Data*, 1–15.
- Goopio, J., & Cheung, C. (2021). The mooc dropout phenomenon and retention strategies. *Journal of Teaching in Travel & Tourism*, 21(2), 177–197.
- Günther, C. W., & Rozinat, A. (2012). Disco: Discover your processes. *BPM (Demos)*, 940(1), 40–44.
- Haar, C. V., & Bell, A. (2020). Analysis of repeat learners in computer science moocs. *2020 IEEE Learning With MOOCS (LWMOOCS)*, 4–7.
- Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Comput. Educ.*, 90, 36–53. <https://doi.org/10.1016/J.COMPEDU.2015.09.005>
- Hosen, M., Ogbeibu, S., Giridharan, B., Cham, T. H., Lim, W. M., & Paul, J. (2021). Individual motivation and social media influence on student knowledge sharing and learning performance: Evidence from an emerging economy. *Comput. Educ.*, 172, 104262. <https://doi.org/10.1016/J.COMPEDU.2021.104262>
- Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Comput. Intell. Neurosci.*, 2018, 6347186:1–6347186:21. <https://doi.org/10.1155/2018/6347186>
- Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students’ quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior*, 92, 496–506.
- Kabudi, T., Pappas, I. O., & Olsen, D. H. (2021). Ai-enabled adaptive learning systems: A systematic mapping of the literature. *Comput. Educ. Artif. Intell.*, 2, 100017. <https://doi.org/10.1016/J.CAEAI.2021.100017>
- Kang, M., & Im, T. (2013). Factors of learner–instructor interaction which predict perceived learning outcomes in online learning environment. *Journal of Computer Assisted Learning*, 29(3), 292–301.
- Kendall, M. G. (1948). Rank correlation methods.
- Khalil, M., & Ebner, M. (2017). Clustering patterns of engagement in massive open online courses (moocs): The use of learning analytics to reveal student categories. *Journal of computing in higher education*, 29, 114–132.
- King, R. B., & Trinidad, J. E. (2021). Growth mindset predicts achievement only among rich students: Examining the interplay between mindset and socioeconomic status. *Social Psychology of Education*, 24(3), 635–652.
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In D. Suthers & K. Verbert (Eds.), *Third conference on learning analytics and knowledge, LAK '13, leuven, belgium, april 8-12, 2013* (pp. 170–179). ACM. <https://doi.org/10.1145/2460296.2460330>

- Krishna, K., & Murty, M. N. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3), 433–439.
- Kurhila, J., Miettinen, M., Nokelainen, P., & Tirri, H. (2004). The role of the learning platform in student-centred e-learning. *IEEE International Conference on Advanced Learning Technologies, 2004. Proceedings.*, 540–544.
- Labrović, J. A., Petrović, N., Anđelković, J., & Meršnik, M. (2023). Patterns of behavioral engagement in an online english language course: Cluster analysis. *Journal of Computing in Higher Education*, 1–26.
- Ledermüller, K., & Fallmann, I. (2017). Predicting learning success in online learning environments: Self-regulated learning, prior knowledge and repetition. *Zeitschrift für Hochschulentwicklung*, 12(1), 79–99.
- Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, 59, 593–618.
- Li, Q., & Baker, R. (2018). The different relationships between engagement and outcomes across participant subgroups in massive open online courses. *Comput. Educ.*, 127, 41–65. <https://doi.org/10.1016/J.COMPEDU.2018.08.005>
- Lin, C.-H., Zheng, B., & Zhang, Y. (2017). Interactions and learning outcomes in online language courses. *British Journal of Educational Technology*, 48(3), 730–748.
- Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an open learner model. *User Modeling and User-Adapted Interaction*, 27, 55–88.
- Lu, X., Wang, S., Huang, J., Chen, W., & Yan, Z. (2017). What decides the dropout in moocs? *Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27-30, 2017, Proceedings 22*, 316–327.
- Malm, E., & Defranco, J. F. (2012). Toward a student-centered measure of learning management system utilization. *Journal of Educational Technology Systems*, 40(4), 401–413.
- Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, 245–259.
- Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. *Proceedings of the 9th international conference on learning analytics & knowledge*, 461–470.
- Monaghan, P., & Rowson, C. (2008). The effect of repetition and similarity on sequence learning. *Memory & cognition*, 36, 1509–1514.
- Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69–84.
- Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86–97.
- Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3), 678–693.

- Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2), 126–135.
- Pope, J. (1972). An investigation of the accuracy of virtual population analysis using cohort analysis. *ICNAF Research Bulletin*, 9(10), 65–74.
- Ramesh, A., Goldwasser, D., Huang, B., III, H. D., & Getoor, L. (2014). Learning latent engagement patterns of students in online courses. In C. E. Brodley & P. Stone (Eds.), *Proceedings of the twenty-eighth AAAI conference on artificial intelligence, july 27 -31, 2014, québec city, québec, canada* (pp. 1272–1278). AAAI Press. <https://doi.org/10.1609/AAAI.V28I1.8920>
- Samuelsen, J., Chen, W., & Wasson, B. (2019). Integrating multiple data sources for learning analytics—review of literature. *Research and Practice in Technology Enhanced Learning*, 14, 1–20.
- Santoso, H. B., Schrepp, M., Isal, R., Utomo, A. Y., & Priyogi, B. (2016). Measuring user experience of the student-centered e-learning environment. *Journal of Educators Online*, 13(1), 58–79.
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3), 1–21.
- Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, 107, 105512.
- Sen, P. K. (1968). Estimates of the regression coefficient based on kendall’s tau. *Journal of the American statistical association*, 63(324), 1379–1389.
- Sennhenn-Kirchner, S., Goerlich, Y., Kirchner, B., Notbohm, M., Schiekirka, S., Simmenroth, A., & Raupach, T. (2018). The effect of repeated testing vs repeated practice on skills learning in undergraduate dental education. *European Journal of Dental Education*, 22(1), e42–e47.
- Serban, C., & Vescan, A. (2019). Advances in designing a student-centered learning process using cutting-edge methods, tools, and artificial intelligence: An e-learning platform. *Proceedings of the 1st ACM SIGSOFT International workshop on education through advanced software engineering and artificial intelligence*, 39–45.
- Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, 747–748.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400.
- Song, M., Günther, C. W., & Van der Aalst, W. M. (2009). Trace clustering in process mining. *Business Process Management Workshops: BPM 2008 International Workshops, Milano, Italy, September 1-4, 2008. Revised Papers 6*, 109–120.
- Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: materials science and engineering*, 336, 012017.

- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes mathematicae*, 12(85), 173.
- Van den Boom, G., Paas, F., & Van Merriënboer, J. J. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and instruction*, 17(5), 532–548.
- Van Der Aalst, W. (2012). Process mining. *Communications of the ACM*, 55(8), 76–83.
- van Alten, D. C., Phielix, C., Janssen, J., & Kester, L. (2020). Self-regulated learning support in flipped learning videos enhances learning outcomes. *Computers & Education*, 158, 104000.
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in human behavior*, 89, 98–110.
- Vincent, J.-L., Rello, J., Marshall, J., Silva, E., Anzueto, A., Martin, C. D., Moreno, R., Lipman, J., Gomersall, C., Sakr, Y., et al. (2009). International study of the prevalence and outcomes of infection in intensive care units. *Jama*, 302(21), 2323–2329.
- Wen, M., Yang, D., & Rosé, C. P. (2014). Linguistic reflections of student engagement in massive open online courses. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, & A. Oh (Eds.), *Proceedings of the eighth international conference on weblogs and social media, ICWSM 2014, ann arbor, michigan, usa, june 1-4, 2014*. The AAAI Press.
- Wong, J., Baars, M., Davis, D., der Zee, T. V., Houben, G., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and moocs: A systematic review. *Int. J. Hum. Comput. Interact.*, 35(4-5), 356–373. <https://doi.org/10.1080/10447318.2018.1543084>
- Wright, M. C., McKay, T., Hershock, C., Miller, K., & Tritz, J. (2014). Better than expected: Using learning analytics to promote student success in gateway science. *Change: The Magazine of Higher Learning*, 46(1), 28–34.
- You, J. W. (2016). Identifying significant indicators using lms data to predict course achievement in online learning. *The Internet and Higher Education*, 29, 23–30.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice*, 41(2), 64–70.