Sarró-Oláh & Fodor *Research and Practice in Technology Enhanced Learning* (2027) 22:31
https://doi.org/10.58459/rptel.2027.22031

Research and Practice in Technology Enhanced Learning

RESEARCH **Free and Open Access**

Modelling students' behaviour: Mining and clustering digital learning paths to identify at-risk students

Bernadett Sarró-Oláh[1] * and Szabina Fodor [2]

*Correspondence: bernadett.sarro-olah@uni-corvinus.hu
Department of Computer Science, Corvinus University of Budapest, Fővám tér 13-15, Budapest, 1093, Hungary
Full list of author information is available at the end of the article

Abstract

The increasing adoption of digital learning environments has led to the generation of rich behavioural data from student interactions in Learning Management Systems. This study aims to identify distinct learning behaviour patterns and examine their association with course failure. Using Dynamic Time Warping, students are clustered based on their activity trajectories, followed by process mining to visualise differences in their learning pathways. The results reveal five learning strategy clusters, two showing a higher incidence of course failure. In the second phase, a Generalized Additive Model is applied to identify key behavioural characteristics of students who failed the course based on the derived process models. The findings highlight that execution-focused or low engagement patterns are more prevalent among unsuccessful students, while sustained, high individual commitment is associated with successful outcomes. Furthermore, the process models enable the identification of critical time periods and learning pathways characteristic of failure. By integrating trajectory-based clustering with a process-centric approach, the study provides a detailed characterization of behaviours associated with course failure and offers actionable insights for improving learning outcomes.

Keywords: educational data mining, LMS, process mining, dynamic time warping, generalised additive model, course failure detection

Introduction

Understanding and predicting student success in digital learning environments has become a central concern in educational research. Learning Management Systems (LMS) such as Moodle generate detailed event logs, referred to as digital traces (Hinds & Joinson, 2019), which capture students' interactions with learning materials, assessments, and communication tools. These traces provide an objective source of behavioural data, enabling the identification of engagement patterns, learning strategies, and risk factors for underperformance (Papamitsiou & Economides, 2014; Zaharuddin et al., 2024). Prior research has shown that usage patterns, such as access frequency and time management, can serve as predictors of student achievement (Bhutoria, 2022; X. Chen et al., 2022).

Self-regulated learning (SRL) offers a valuable theoretical framework for interpreting such behaviour, describing how learners plan, monitor, and reflect on their learning processes (Zimmerman, 2002). However, understanding how these processes manifest in digital environments requires analytical approaches capable of capturing behavioural patterns in trace data. Educational data mining (EDM) has increasingly been used to uncover such patterns through predictive and exploratory techniques, including clustering, which enables the identification of latent learner profiles (Lin et al., 2023; Wang, 2025). Despite these advances, EDM approaches often provide limited insight into the structure and dynamics of learning processes. To address this limitation, temporal analysis techniques have been applied to capture similarities in time-dependent behavioural patterns (F. Chen & Cui, 2020; Saint et al., 2020). However, while these methods account for timing and sequencing, they do not explicitly model the underlying process structure of learning behaviour. In contrast, process mining (PM) enables the discovery and analysis of sequential patterns, such as learning pathways, offering a process-oriented perspective on learner behaviour (Aalst et al., 2009; Van Der Aalst, 2016). By reconstructing sequences of activities, PM facilitates a more detailed understanding of behavioural processes, although the resulting models may be complex and difficult to interpret (Wijnhoven et al., 2023). While these approaches provide valuable insights, their isolated application limits their explanatory power. Recent studies suggest that combining PM outputs with traditional data analysis can improve predictive performance (Lin et al., 2023; Semler et al., 2025), while integrating temporal and process perspectives enables a more comprehensive understanding of learning behaviour. Despite these developments, several gaps remain. First, limited research integrates temporal and process-oriented analyses within a unified framework. Second, clustering approaches are rarely combined with temporal and process-based modelling to identify behavioural profiles linked to risk. Third, relatively few studies examine the evolution of learning behaviour across an entire semester. To address these gaps, the following research questions are formulated:

RQ1: Which temporal and process-oriented patterns of students' digital learning behaviour can be systematically discovered from LMS event log data?

RQ2: How can these patterns be used to identify critical learning trajectories and time periods associated with academic underperformance or course failure throughout a semester?

The contribution of this study is an integrated framework that links traditional data analysis, temporal dynamics, process structure, and explainable modelling to examine when and how declining engagement may relate to failure risk in LMS-based learning. By combining these perspectives, the study provides indicative insights into critical periods and behavioural patterns that may remain less visible using traditional approaches.

Background

Digital learning environments and self-regulated learning

In contemporary higher education, digital tools play an increasingly important role in supporting the learning process, the trend that was significantly accelerated by the COVID-19 pandemic (Adigun et al., 2025; Ngah et al., 2022). LMS provides structured platforms for delivering instructional content, managing assessments, and facilitating communication (AlQaheri & Panda, 2022; Gasevic et al., 2017). In addition, a key advantage of these platforms is their ability to record user interactions in the form of detailed event logs. These systematically recorded log events constitute a specific, education-related form of digital trace data and are widely used as input in many studies (Bessadok et al., 2023; Fan et al., 2021; E. Real & Pimentel, 2025). Such log data can be interpreted in multiple ways, depending on the analytical perspective or level of aggregation. Therefore, the reading and conceptualisation of digital trace data (Pentland et al., 2020) is crucial.

Interpreting these behavioural traces requires an appropriate theoretical framework linking observable actions to underlying learning processes (Arizmendi et al., 2022). One of the most relevant perspectives in this regard is SRL, which refers to learners' ability to actively plan, monitor, and manage their learning processes in order to achieve their goals, and has been widely conceptualised as a cyclical process consisting of three phases: forethought, performance, and self-reflection (Zimmerman, 2002). Digital learning environments both support and challenge SRL. While they promote flexibility and autonomy (Shirk, 2020), allowing learners to design their own learning strategies (Fan et al., 2021), they also place greater responsibility on learners to manage their engagement and motivation. The results of Jin et al. support this, indicating that AI tools facilitate the cognitive, metacognitive, and behavioural aspects of SRL, while their contribution to motivational regulation is more limited (Jin et al., 2023). A closely related concept is self-directed learning, which also emphasises learner autonomy but places greater responsibility on learners to initiate and manage their own learning (Garrison, 1997). Although these concepts are closely related and often discussed together in the literature, the present study adopts SRL as the overarching framework.

Educational data mining and learning analytics

The increasing availability of digital trace data has enabled the application of data-driven approaches in education. Two closely related research fields addressing this challenge are

Educational Data Mining (EDM) and Learning Analytics (LA), both of which rely on analysing large volumes of educational data to understand learner behaviour and improve learning. Despite this shared focus, EDM traditionally emphasises the development and application of computational and statistical methods to identify patterns in educational data, while LA places greater emphasis on understanding and optimising learning processes within educational contexts (Baker & Siemens, 2014).

In this regard, unsupervised methods such as clustering are commonly used to group learners based on behavioural similarities without predefined labels, enabling the identification of similar interaction patterns in LMS (Bey & Champagnat, 2022; Bogarín et al., 2014). Moreover, such behavioural patterns can reveal differences in study strategies, engagement, and SRL behaviours, helping to explain variations in learning outcomes across groups. Prior research shows that higher activity levels are generally associated with better performance (Cenka et al., 2022), while different behavioural patterns may lead to similar outcomes, suggesting the presence of multiple effective learning pathways (Bessadok et al., 2023). This is further supported by findings that conceptualize learning pathways as dynamic complex systems (Ortiz-Vilchis & Ramirez-Arellano, 2023), where properties such as equifinality allow different learning paths to lead to similar outcomes, indicating that no single optimal pathway exists. In addition, the growing availability of digital trace data highlights the importance of their proper analysis and interpretation. Leveraging LA and AI-based approaches, particularly explainable AI, enhances insight generation by revealing key factors influencing learner behaviour and supporting more targeted interventions (Kalita et al., 2025; Palanci et al., 2024).

Temporal and process-oriented analysis of learning behaviour

In recent years, growing attention has been devoted to *temporal patterns of learner behaviour*. Event logs inherently capture the sequential nature of learning processes, allowing the analysis of not only what learners do, but also when and in what order activities occur. Recent studies show that incorporating temporal behavioural patterns into analysis improves performance prediction compared to static features (F. Chen & Cui, 2020) and reflects differences in SRL-related activity sequences (Saint et al., 2020). Building on this, learning behaviour can be understood as evolving *learning trajectories*, where both the timing and structure of activities shape outcomes. Previous research has shown that behavioural trajectories and engagement dynamics are closely associated with learning outcomes (Mukala, 2025) and more successful learners tend to follow structured and goal-oriented sequences, whereas less successful students exhibit more fragmented patterns (Li et al., 2023). At the same time, different learning pathways may lead to similar results (Ortiz-Vilchis & Ramirez-Arellano, 2023). Distinctions can also be identified through learner profiles derived from interaction sequences (Maldonado-Mahauad et al., 2018), which can be captured using approaches such as Transition Network Analysis to model temporal transitions between learning activities (Saqr et al., 2025). Such patterns have also been used to inform personalised learning interventions (Hachicha et al., 2021), with related work demonstrates their implementation through adaptive pathways and gamification-based approaches tailored to individual learners (Xiao & Hew, 2024). In addition, visualisation techniques enhance the interpretability of complex behavioural patterns and support the identification of performance-related trajectories (Feng et al., 2022).

While temporal analysis provides valuable insights, learning behaviour can be more comprehensively understood when conceptualized as a process. PM facilitates the modelling of learning behaviour as structured pathways based on the sequencing and interconnection of activities (Aalst et al., 2009; Van Der Aalst, 2016). Commonly used discovery algorithms, including Alpha Miner (AM), Inductive Miner (IM), and Heuristic Miner (HM), differ in their ability to handle noise and model behavioural complexity (Van Der Aalst et al., 2004), with IM and HM being more suitable for educational data. While IM produces more robust and structurally sound models (Cerezo et al., 2020; Leemans et al., 2014; Romero et al., 2008), HM captures more flexible behavioural patterns by relaxing strict event ordering (Weijters et al., 2006). Application of educational process mining (EPM) show that students often deviate from intended learning paths, and such deviations, along with structural properties of learning processes, are associated with performance differences (AlQaheri & Panda, 2022; Cenka et al., 2022; Martínez-Carrascal et al., 2024). Process-based analyses further reveal that high-performing students tend to exhibit more structured and coherent learning processes, while less successful learners show more irregular patterns (Hartmann et al., 2022; E. M. Real et al., 2020). To address the heterogeneity of learner behaviour, clustering is frequently combined with process mining, enabling the identification of groups with similar behavioural strategies and improving the interpretability of process models (Bey & Champagnat, 2022; Bogarín et al., 2014). Such integrated approaches have been shown to reveal distinct learner profiles and provide actionable insights into learning pathways (Chanifah et al., 2021; Patel et al., 2017).

From Learning Behaviour to Outcomes and Risk

The likelihood of academic attrition increases with course failure (Ajjawi et al., 2020), highlighting the importance of early identification of at-risk students. LA plays a key role in this process, as behavioural data can reveal learning strategies associated with performance differences (Gasevic et al., 2017). Prior research shows that academic performance is shaped by both individual and instructional factors, including motivation, study habits, and teaching practices (Saha et al., 2022). One of the most critical determinants of success is student engagement, as uneven or poorly timed participation across the learning cycle is associated with a higher likelihood of underperformance and dropout (Hernández-García et al., 2024). This is further shaped by students' interpretations of failure and their emotional responses, which influence subsequent engagement (Ajjawi et al., 2020). Given its temporal nature, engagement can be captured through time-series analysis of LMS data, which measures activity frequency, consistency, and sequencing (F. Chen & Cui, 2020). Building on this, behavioural trajectories can reveal how attrition emerges gradually from interaction patterns and deviations over time (Brochenin et al., 2017). Extending this perspective, recent multimodal approaches strengthen early risk detection by capturing complex temporal and behavioural patterns (Sridharan & Sharagaharajan Akilashri, 2025). From a structural viewpoint, process-oriented analyses suggest that performance depends not only on activity levels, but also on how learning activities are organized into coherent sequences (Juhaňák et al., 2019). Accordingly, modelling learning pathways further improves performance prediction (Umer et al., 2017), with recent work extending predictive models by integrating PM with deep learning to incorporate sequential behavioural information (Helal et al., 2025). From SRL perspective, differences between successful and unsuccessful learners are evident in learning structure and regulation, with high-performing students showing more coherent and strategic regulation cycles (Cerezo et al., 2020). These insights can also be translated into practice through LA dashboards that support self-regulation and enable early identification of at-risk students (Romero et al., 2008; Susnjak et al., 2022).

Method

Course and participants

The selection of the course and participant group is a critical aspect of the study, as it shapes the validity and interpretability of the results. The study focuses on the course Fundamentals of Artificial Intelligence, a STEM subject requiring both theoretical understanding and practical competencies, based on data from the spring semester of the 2024/2025 academic year (17 February–23 June 2025) at Corvinus University of Budapest. All 103 participants were third-year undergraduate Business Informatics students with prior experience in practice-oriented courses and Moodle use. The course was delivered through weekly in-person sessions requiring continuous individual preparation. A subgroup (n = 9) followed an individual study schedule due to studies abroad, relying exclusively on Moodle materials. Course materials, including theory, Python tasks, and supplementary resources, were uploaded on the day of each class, with solutions released the following Friday to support self-evaluation. Expected learning behaviour involved active participation, independent task completion, and self-assessment; however, Moodle use was optional and not formally monitored. Moodle was nevertheless required for submitting three individual assignments and completing two mid-term tests, forming a 100-point grading system. The exact deadlines were specified in the course syllabus. Students who did not meet the requirements or were dissatisfied with their performance could take a final exam. For the purposes of this study, students who did not obtain a final (offered) grade were classified as non-completers. By the end of the semester, 97 students successfully met the course requirements, while 6 were classified as non-completers.

Data collection and preparation

The analysis is based on event log data extracted from Moodle. The dataset includes attributes such as timestamp, name, user ID, event context, component, event name, and description, enabling each activity to be linked to individual students. Although action durations cannot be directly measured, timestamps provide insight into step-by-step behaviour. The attributes component, event name, and event context were used to define activity types; based on their combination, custom categories and subcategories were established (see Figure 9 in the Appendix). Data preprocessing involved removing erroneous, duplicate, and irrelevant entries. The final dataset comprised 25 484 events.

Dynamic time warping

This research identifies typical digital learning paths using Dynamic Time Warping (DTW), a time-dependent method that captures recurring behavioural patterns and enables the analysis of non-linear relationships, even between sequences of different lengths (Ghanem, 2013). As a result, this method generates homogeneous groups through an iteratively calculated central time series (Petitjean et al., 2011), with clustering performed on event-level LMS log data enriched with activity category and subcategory labels, without prior aggregation. The number of clusters was determined by jointly considering the Elbow method (Schubert, 2023) and DTW-specific indices based on medoids, defined as the most central instance within a cluster (Koschke & Steinbeck, 2020; Petitjean et al., 2011). Cluster quality was evaluated using the Calinski-Harabasz (CH) and Davies-Bouldin (DB) indices to assess separation and compactness (López-Rubio et al., 2018; Ngo & Phan, 2025). The final number of clusters was determined by comparing these metrics and aligning them with the research objectives, favouring a smaller number of sufficiently distinct clusters when no clear optimum was indicated. The clustering was performed using the TimeSeriesKMeans algorithm (Tavenard et al., 2020), with settings ensuring reproducibility of the results.

Process mining

PM enables analysis of sequential learning behaviour, with process discovery performed using directly-follows graphs (DFGs) and retaining connections with a minimum frequency of 20 to highlight dominant patterns (Leemans et al., 2018). Petri-nets were also constructed using the AM, IM, and HM algorithms to represent processes in greater detail. This allows explicit analysis of places and transitions (Lassen & Van Der Aalst, 2009), and supports advanced metrics for model comparison. Each process discovery algorithm requires specific configuration settings; therefore, a uniform noise threshold of zero was applied across all models to ensure methodological consistency. While the HM offers extensive flexibility through multiple parameters, this study employed standard settings, dependency_threshold (0.5), and_threshold (0.65), loop_two_threshold (0.5), min_act_count (1), min_dfg_occurrences (1), and frequency-based decoration, to ensure comparability across algorithms (Berti et al., 2023; Van Der Aalst, 2016). Models generated by different algorithms were evaluated through conformance checking with four key metrics: token-based fitness, which measures the alignment between the model and the event log traces (Buijs et al., 2014); precision, which evaluates whether the model avoids allowing behaviour not observed in the log (Buijs et al., 2014); generalisation, which reflects the model's ability to handle unseen but plausible behaviour (Buijs et al., 2014; Van Der Aalst, 2016); and simplicity, which captures the structural complexity and interpretability of the model (Buijs et al., 2014; Vázquez-Barreiros et al., 2015). The selection of final model is based on the average of these metrics, assigning equal weight to each to ensure a balanced evaluation across all dimensions.

Generalized additive model

To support further analysis, a Generalized Additive Model (GAM) (Hastie & Tibshirani, 1986, 2014) was applied, enabling the exploration of nonlinear relationships and the temporal effects of predictors on course completion outcomes. Its application is particularly appropriate, as student behaviour may change over the semester, influencing academic outcomes. Given the binary nature of the dependent variable, a logistic GAM was applied using the pyGAM module (Chang et al., 2021; Servén et al., 2025; Zhang et al., 2018). The analysis was based on event log data enriched with cluster labels, course completion status, and structured by temporal segments. To address class imbalance in the data, stratified splitting, class weighting, and threshold optimisation were applied. Prior to modelling, multicollinearity was assessed using the Variance Inflation Factor (VIF) (Bondaug & Tubo, 2024). Parameter tuning focused on selecting the optimal lambda values using grid search (0.5, 0.6, 0.9, 1), balancing model smoothness and overfitting (S. N. Wood, 2025). Model performance was evaluated using multiple metrics, including Effective Degrees of Freedom (DoF), log-likelihood, and the corrected Akaike Information Criterion (AICc), complemented by pseudo-R^2 to assess explanatory power (Detmer et al., 2025; Hemmert et al., 2018; S. Wood, 2001). While the best-performing model was selected based on these criteria, the interpretation of predictor significance was treated with caution due to potential bias introduced by spline-based estimation (S. N. Wood, 2025).

Feature importance analysis

To support the interpretation, the Local Interpretable Model-agnostic Explanations (LIME) technique was applied (Ribeiro et al., 2016). LIME approximates the model locally, enabling instance-level assessment of feature contributions for individual predictions and improving model transparency. Aggregating these explanations reveals general patterns of feature importance, providing an additional layer of validation for the GAM results. The same training and testing datasets were used, with classification mode applied to match the binary outcome, consistent with the GAM setup.

Findings

Distribution and intensity of activity usage

Following the data preparation, descriptive statistics were generated to summarise students' learning activities. The recorded events were grouped into eight main activity categories (see Figure 9 in the Appendix). Assignment activities, exam activities, tasks, and general information were used by all students at least once during the semester. In contrast, two students did not access the available presentations, while multimedia content and extra content engaged 85 and 83 students, respectively. Attendance-related activities were observed for 35 students; however, as this feature was available but not part of the course design and showed low activity level, it was excluded from further analysis. In terms of total activity counts, assignment and exam activities were the most frequently occurring categories (see Table 1).

Table 1

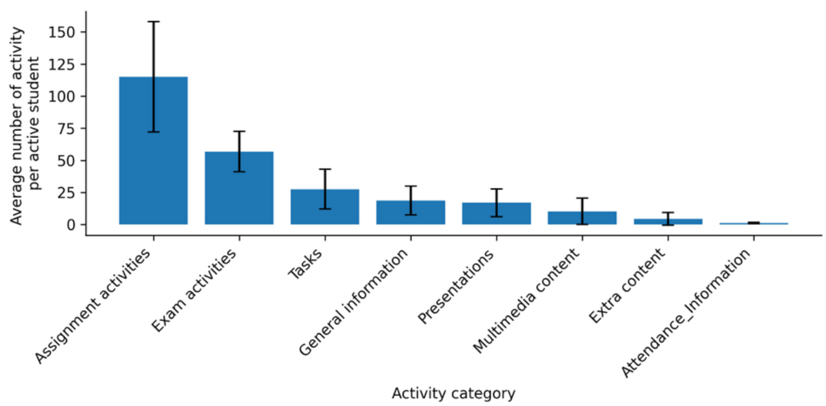
Number of users across the activity categories throughout the semester (prepared by the authors)

Activity_category	Number of users	Total number of activities
Assignment activities	103	11856
Exam activities	103	5857
Tasks	103	2837
General information	103	1923
Presentations	101	1717
Multimedia content	85	876
Extra content	83	375
Attendance	35	43

Activity categories at the student level (see Figure 1) are consistent with the results presented previously: assignment and exam activities exhibit the highest average usage, accompanied by considerable variability. Other activity types occur less frequently on average, although their usage still varies substantially across students.

Figure 1

Average usage intensity across activity categories per student (prepared by the authors)

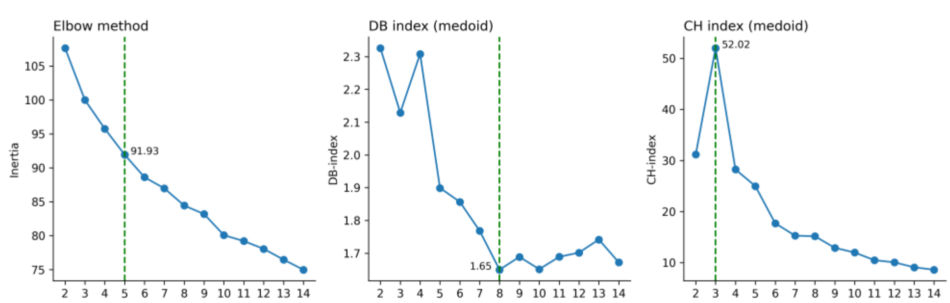


Digital learning pattern recognition

Following the initial data overview, clustering analysis is performed using event-level data enriched with activity categories and subcategories. Clusters were formed using DTW-based clustering. As an unsupervised approach, selecting the optimal number of clusters was supported by the Elbow method, DB index, and CH index (see Figure 2).

Figure 2

Metrics of clustering (prepared by the authors)

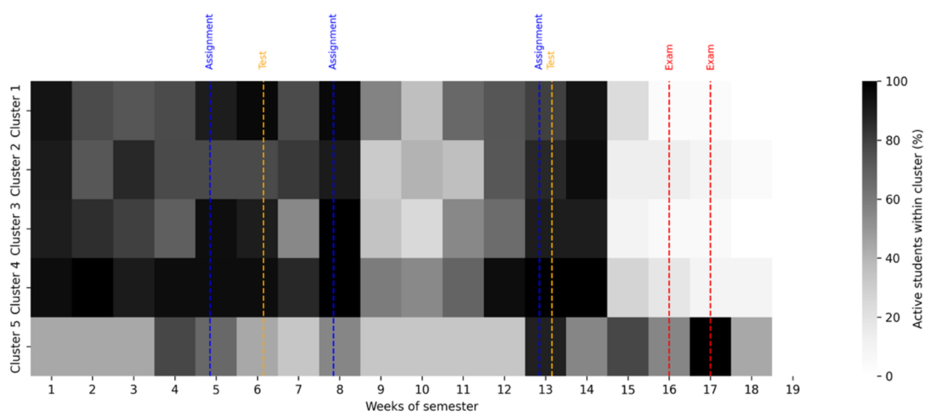


The metrics suggested different solutions: five clusters (inertia = 91.93), eight clusters (DB = 1.65), and three clusters (CH = 52.02), reflecting their differing evaluation criteria. The three-cluster solution was rejected due to poor separation (high DB value), while the eight-cluster solution was not supported by the CH index. Therefore, five clusters were selected as a balanced compromise, providing acceptable metric values and a stable, interpretable distribution of instances.

Cluster sizes vary, with Cluster 1 (n=30), Clusters 2 and 4 (n=22 each), Cluster 3 (n=20), and Cluster 5 (n=9). In general, student activity is largely driven by course-related assessments (tests, assignments, exams). To better capture differences between clusters (see Figure 3**)**, these days are excluded from interpretation but retained in the overall analysis. Activity was defined at the weekly level, with students considered active if they performed at least one action, allowing proportion-based comparisons across clusters of different sizes.

Figure 3

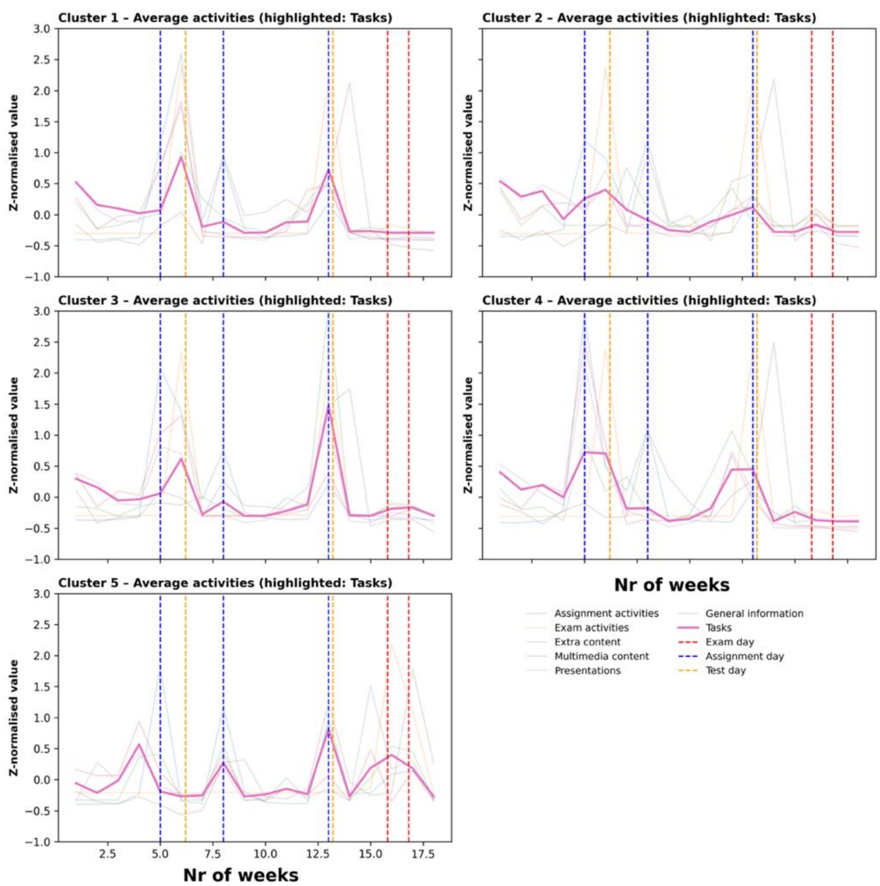
Weekly learning-related activity patterns across clusters, excluding activities generated on test, assignment, and exam days (prepared by the authors)



Despite excluding assessment days, activity peaks remain visible around those weeks in Clusters 1–4. These clusters also exhibit higher activity at the beginning of the semester, followed by a gradual decline, with Cluster 4 maintaining relatively higher engagement in the later phases. In contrast, Cluster 5 shows increasing activity toward the end of the semester. To better interpret differences in student participation beyond absolute activity levels, activity was analysed using cluster-level standardised measures. Standardisation (z-scores) was applied to ensure comparability across activity categories with different frequencies (see Figure 4). Particular emphasis was placed on the non-mandatory Tasks category, as it reflects hands-on programming engagement. *Cluster 1* exhibits a strong initial engagement, which declines by the end of the second week. The activity is increasing approximately one week before assessments in this group, followed by minimal engagement during the exam period. *Cluster 2* shows lower overall activity. Similar to Cluster 1, initial engagement declines quickly, with moderate increases around assessments and less pronounced task-related peaks, suggesting lower engagement in practical digital activities. *Cluster 3* shows slightly above-average initial engagement and closely follows assessment schedules, with activity peaks occurring shortly before deadlines. In the second half of the semester, task-related engagement is relatively more pronounced. Only a minor increase is observed during the exam period. *Cluster 4* demonstrates a moderately elevated start, with earlier preparation beginning 1.5–2 weeks before assessments. A decline occurs between tests, and activity remains low during the exam period, with only a slight increase beforehand. *Cluster 5* shows a delayed start, with higher activity emerging around weeks 3–4. Engagement is limited before the first test but increases toward later assessments. Unlike other clusters, activity peaks during the exam period, with the longest preparation phase (approximately 2.5–3 weeks). Overall, the clusters exhibit distinct, though partly subtle, behavioural patterns in their practical learning processes.

Figure 4

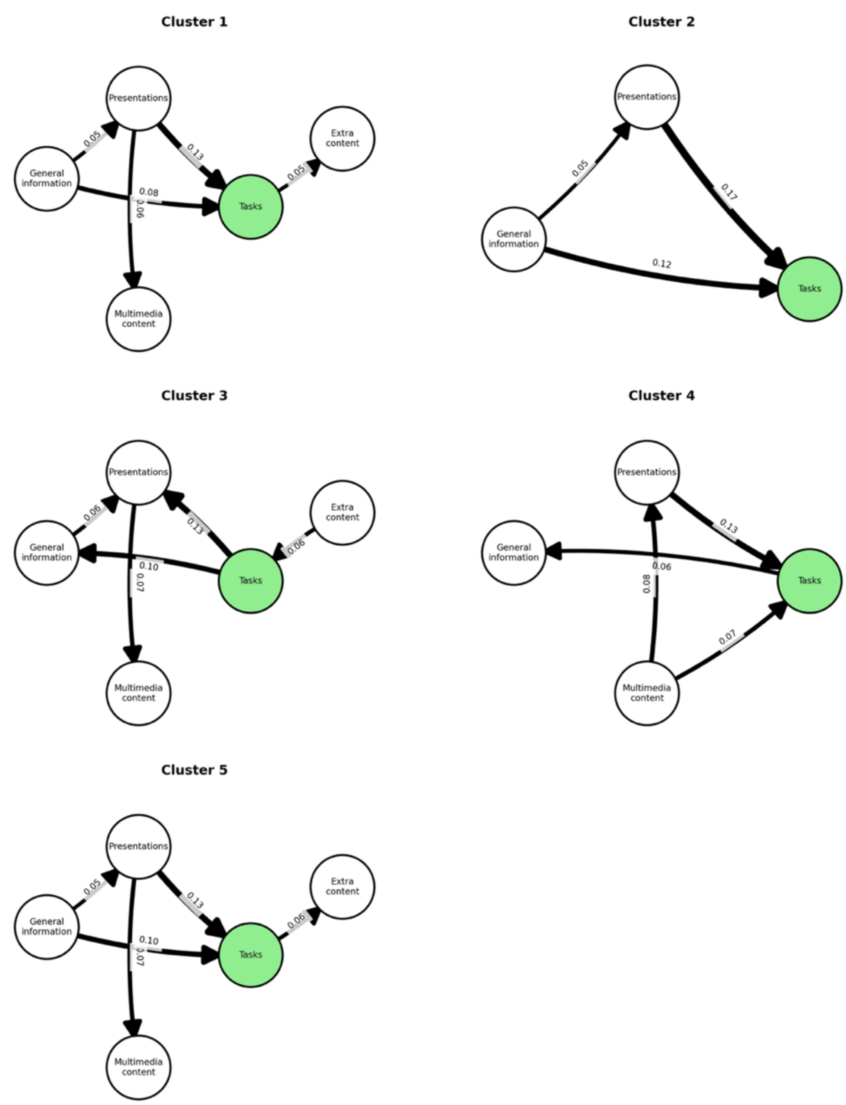
Digital learning activities of the clusters; highlighted: Tasks (prepared by the authors)



As a process model, DFGs (see Figure 10 in Appendix) are able to illustrate cluster differences based on activity transitions, yet even after frequency filtering, key distinctions remain difficult to identify. To improve interpretability, the analysis focuses on the most frequent transitions within each cluster, where the direction and relative frequency of transitions reveal key differences. To highlight the underlying learning behaviour, only freely chosen activities are considered, excluding assignment, test, and exam-related actions, which, although dominant, are less informative from an SRL perspective (see Figure 5 for details).

Figure 5

Dominant activity transitions across clusters (non-assessment activities; prepared by the authors)



Cluster 1 is characterised by strong convergence towards Tasks, primarily driven by transitions from Presentations and General information, indicating a structured progression from content consumption to practice. Secondary transitions to Multimedia and Extra content suggest some exploratory behaviour. Overall, this cluster reflects a task-oriented yet relatively diverse learning pattern. *Cluster 2* is characterised by a simpler, less diverse

transition structure, with activity primarily directed from General information and Presentations to Tasks. This reflects a more linear progression with limited exploration and few alternative pathways, indicating a minimalistic, task-oriented approach relying mainly on core course materials. *Cluster 3* exhibits a structure similar to Cluster 1, but with reversed transition directions. Dominant transitions originate from Tasks and lead to Presentations and General information, indicating a more iterative, feedback-oriented pattern rather than a linear progression. *Cluster 4* exhibits a structured transition pattern, with Multimedia content acting as a central link between Presentations and Tasks. The dominant transitions indicate a sequential flow from content consumption through multimedia-supported materials to task engagement. Compared to other clusters, this reflects more integrated use of diverse learning resources. *Cluster 5* shows a transition structure similar to Cluster 1, with dominant flows from Presentations and General information to Tasks. In addition, transitions involving Multimedia and Extra content are more pronounced, indicating a stronger engagement with supplementary materials alongside the content-to-task progression.

Patterns of course non-completion

According to the interpretation of course non-completion applied in this study, a total of six students did not complete the course. These students were distributed across three clusters: three in Cluster 1 (representing 10% of that cluster), two in Cluster 2 (9% of that cluster), and one in Cluster 3 (5% of that cluster). In contrast, all students in Clusters 4 and 5 successfully completed the course. Despite the presence of non-completing students in the first three clusters, their average performance remained relatively high, whereas Cluster 4 and 5 achieved the highest average score overall. Examining the learning patterns of non-completing students is therefore particularly valuable. As student activity levels varied across the semester, distinct periods were identified to capture these differences:

- the start period (17-02-2025 to 16-03-2025)
- before the first assignment (17-03-2025)
- before the first test (24-03-2025 to 25-03-2025)
- before the last assignment and test (28-04-2025 to 13-05-2025)
- before the exam period (15-05-2025 to 01-06-2025)
- the exam period (02-06-2025 to 12-06-2025)

Differences in sequence length do not affect the analysis, as it focuses on learning structure rather than activity volume, with periods compared consistently across students. PM was applied at the student-period level, with models generated using multiple discovery algorithms (AM, IM, HM). For each student-period, the best-fitting model was selected, and its simplicity value was used as an indicator of behavioural complexity in subsequent analyses. As simplicity showed a strong inverse relationship with activity volume (r = –0.767), it provides a compact measure of how structured or fragmented learning behaviour is.

Before further analysis, relationships between period-specific variables were assessed. VIF results (1.07–1.22) indicate that multicollinearity is not a concern. Following preprocessing, the dataset exhibited strong class imbalance, with few failure cases. To address this, stratified splitting (30% test), minority class weighting (weight = 3), and threshold optimisation were applied. Threshold selection was based on cross-validation; however, due to the limited number of failures, some folds (10/100) did not converge, reflecting variability across splits. The classification threshold was set to 0.20 to improve at-risk detection, with model selection prioritising recall over overall accuracy, resulting in increased detection at the cost of more false positives. A logistic GAM was then fitted using optimised smoothing parameters. The resulting model showed Effective DoF = 11.73, log-likelihood = −5.21, and AICc = 37.77, with a high pseudo-R^2 (0.880), although this should be interpreted with caution given the limited number of failure cases. The significance of smooth terms and their associated smoothing parameters (lambda) are summarised in Table 2.

Table 2

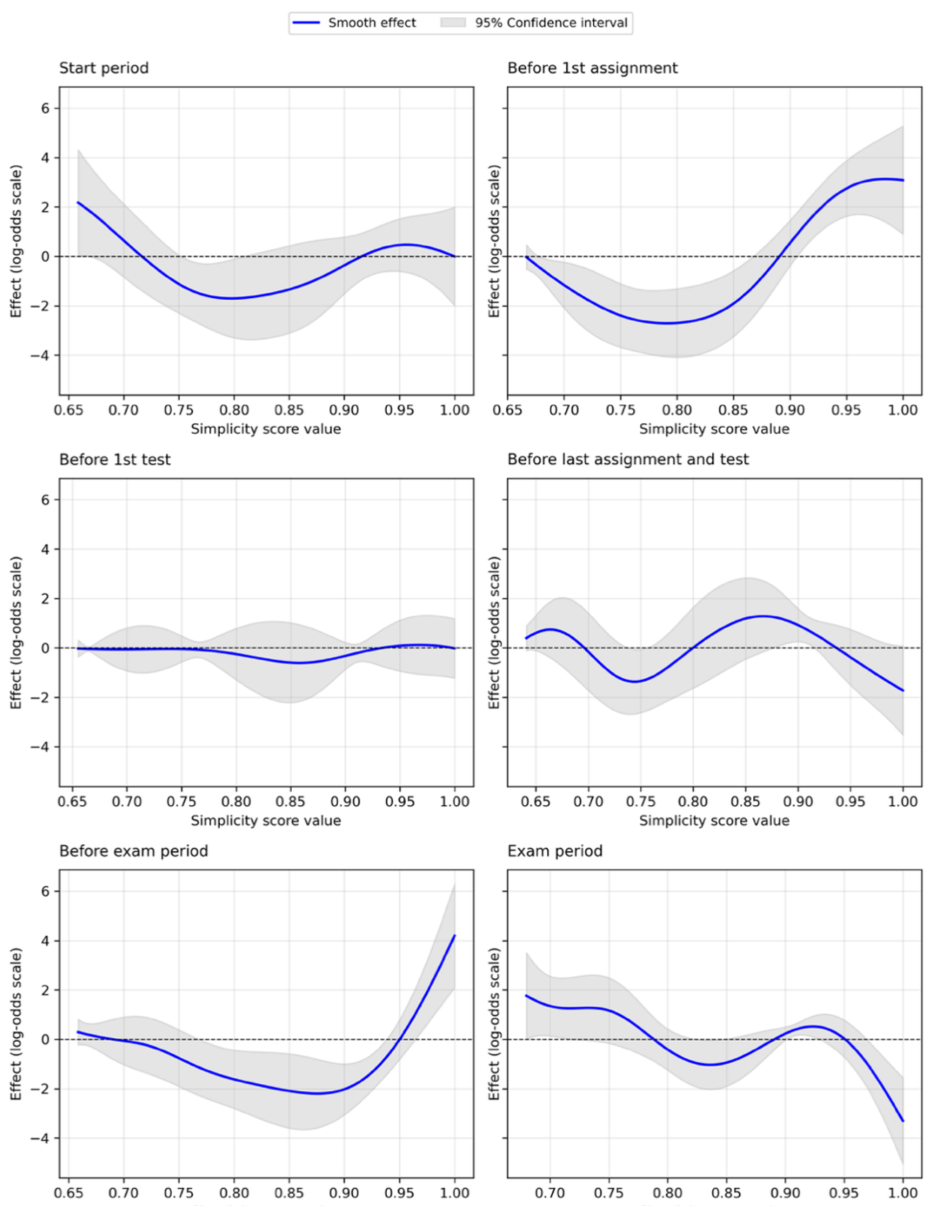
Estimated smoothing parameters (lambda) and significance of the GAM smooth terms. P-values are reported with two decimal precision. Significance levels: * $p < 0.05$, *** $p < 0.001$. (prepared by the authors)

Variable	Lambda	p-value
Start period	0.6	0.03*
Before first assignment	1.0	<0.01***
Before first test	0.5	<0.01***
Before last assignment and test	0.5	0.11
Before exam period	0.9	<0.01***
Exam period	0.5	<0.01***

On the full sample, the model achieved seemingly strong performance (accuracy = 0.951, balanced accuracy = 0.974, precision = 0.545, recall = 1.000, F1 = 0.706, AUC = 1.000). However, given the small number of failure cases (n = 6) and full-sample evaluation, the results, including AUC, should be interpreted cautiously and as exploratory, as they may reflect optimistic performance rather than true generalisation. While all failure cases were identified, this came at the cost of lower precision, indicating false positives. Overall, the findings suggest the model's potential for identifying at-risk students. Figure 6 displays the partial dependence plots from the GAM, illustrating the estimated non-linear effects of predictors on the log-odds of failure, averaged over other variables. The x-axis shows predictor values, while the y-axis represents their effect on failure risk. The solid line indicates the smooth effect, with shaded areas denoting 95% confidence intervals. Values above zero indicate increased risk, while values below zero indicate decreased risk; accordingly, higher simplicity scores (i.e., lower activity levels) are associated with a higher likelihood of failure. The effects vary across time periods, with clear differences in magnitude.

Figure 6

Impact of process simplicity scores on course failure across study periods (prepared by the authors)

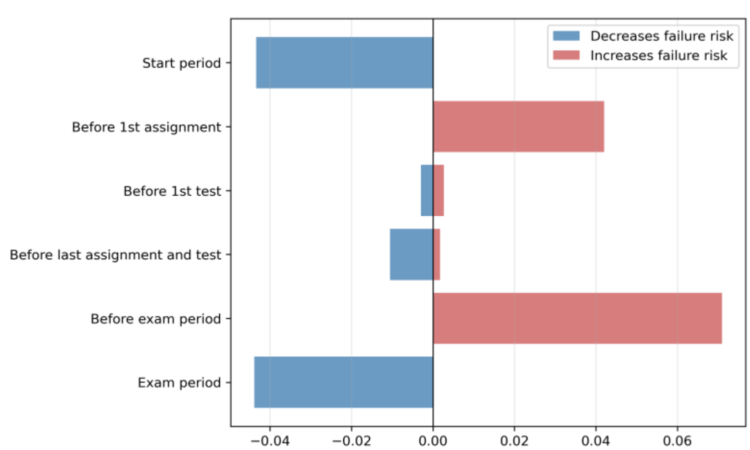


In the start period, higher activity (lower simplicity values around 0.65-0.70) is associated with reduced risk, while risk increases at higher simplicity levels (around 0.90-1.00). A stronger pattern is observed before the first assignment, where the effect increases from approximately −3 (around 0.75-0.80) to approximately +3 at higher simplicity values (around 0.95-1.00). The effect before the first test remains weak (roughly between −1 and 0.5), suggesting limited influence. The period before the last assignment and test shows a non-linear pattern, with the effect decreasing to approximately −1.5 around 0.70-0.75, increasing to approximately +1.5 around 0.85, and declining again at higher simplicity values; however, this effect is not statistically significant. The strongest effect is observed before the exam period, where the effect increases sharply from approximately −2 (around 0.85) to approximately +4 at high simplicity values (around 1.00). During the exam period, the relationship is highly non-linear, with an initial decrease followed by a brief increase, and a sharp decline at the highest simplicity values. Overall, the relationship between simplicity scores and failure risk is non-uniform and phase-dependent, with higher simplicity (lower activity level) associated with increased risk in some stages but more favourable outcomes in others, such as the exam period. A pattern appears to emerge in the start period, suggesting potential early signs of risk, and becomes more pronounced before the exam, where declining engagement may serve as an indicator. Together, these findings suggest that the effects of simplicity evolve over the course.

To further support GAM results, LIME was applied to quantify feature contributions (see Figure 7). Activity in the start and exam periods is associated with reduced failure risk, whereas activity before the first assignment, and especially before the exam period, is associated with increased risk.

Figure 7

Results of the LIME analysis (prepared by the authors)



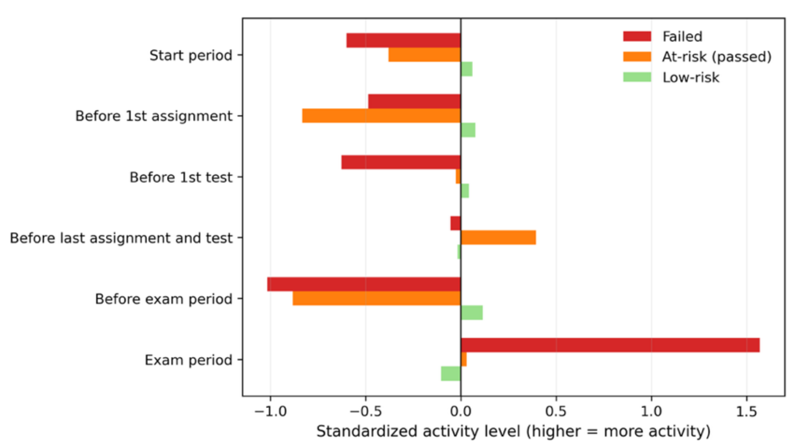
The strongest risk-increasing contribution is observed before the exam, highlighting the impact of declining engagement at this stage. In contrast, effects around the first test and before the last assignment and test are comparatively small. Contributions are clearly separated by direction, indicating strong polarity across periods. In summary, the LIME results support the GAM findings, confirming the key role of reduced late-stage activity and the protective effect of earlier engagement.

Under the selected parameter settings, the model classified 92 students as low-risk (passed), correctly identified all 6 failure cases, and flagged 5 additional students as at-risk despite ultimately passing. Of the five at-risk students, three belonged to Cluster 1, while one student each was assigned to Cluster 3 and Cluster 4. It is informative to examine how these groups of students differ in their behavioural patterns across the observed time periods (see Figure 8).

Failed students and those classified as at-risk but ultimately passing exhibit largely similar behavioural trajectories across most periods, consistently reflecting below-average activity levels. A divergence is observed in the before last assignment and test phase, where the at-risk group demonstrates relatively higher activity. In contrast, low-risk students display an opposing pattern, maintaining near-average or slightly above-average activity throughout the semester, with a decline emerging only during the exam period. Notably, from the start period to before the first assignment, failed students are initially more passive but show increasing activity over time, whereas at-risk (passed) students exhibit the opposite pattern, becoming more passive after a less negative starting point.

Figure 8

Low-risk, at-risk and failing student profiles of student groups across periods (prepared by the authors)



Based on behavioural patterns across the semester, process model characteristics, and outcome distributions (failed, at-risk, and passed), the clusters were interpreted and labelled as follows:

Cluster 1 (*EFDL – Execution-Focused Digital Learners*) is characterized by structured, task-oriented behaviour, with activity concentrated around key course requirements. The pattern reflects a focus on task completion rather than sustained engagement and is associated with increased risk.

Cluster 2 (*LEDL – Low Engagement Digital Learners*) is characterized by limited digital activity throughout the semester, with process models indicating sparse interaction patterns. However, this does not necessarily imply a lack of learning, but rather lower reliance on the digital environment. Compared to Cluster 1, this group shows fewer failed and no at-risk cases, suggesting that students may adopt alternative, less digitally engaged learning strategies.

Cluster 3 (*RSDL – Reactive Strategic Digital Learners*) is characterized by effective use of digital tools, with activity extending beyond mandatory tasks. The transition structure shows bidirectional interactions between tasks and learning materials, indicating iterative, need-driven and reactive engagement concentrated around key requirements. This pattern is associated with predominantly successful outcomes, with only limited at-risk and failure cases.

Cluster 4 (*CDL – Consistent Digital Learners*) exhibits generally stable and sustained engagement throughout the semester, supported by well-structured process models. This cluster is predominantly associated with successful outcomes and low risk, with only occasional at-risk case.

Cluster 5 (*IDL – Independent Digital Learners*) is characterized by relatively low but efficient and self-regulated activity, with process patterns suggesting selective interaction. This cluster is associated with a self-paced learning approach and successful outcomes. Notably, it includes all students following an individual study plan, reinforcing its interpretation as independent learners.

Discussion

Following the analytical logic of the study, the findings are interpreted in relation to the existing literature. In general, student activity is largely concentrated around assessment periods, indicating that engagement is strongly driven by course requirements. This aligns with prior findings showing that LMS usage is often selective rather than comprehensive (Cerezo et al., 2016), shaped by external expectations (Nizam Ismail et al., 2021) and individual preferences (Alseitova et al., 2025). Accordingly, digital learning behaviour reflects goal-oriented and context-dependent engagement rather than uniform resource use.

The inclusion of the temporal dimension proves essential, as engagement and learning strategies evolve throughout the semester. Rather than being static, learning behaviour unfolds as a dynamic process in which patterns emerge, stabilise, or shift over time. Consistent with prior research (Cristea et al., 2025), effective learning is often associated with early and sustained engagement; however, the present findings also indicate that multiple, distinct strategies may lead to successful outcomes, emphasising the validity of diverse learning pathways.

Analysing behavioural trajectories across comparable time periods provides a more nuanced understanding of SRL and engagement, as aggregate patterns may mask substantial heterogeneity in learning behaviour. Clustering has been commonly applied in previous research (Cerezo et al., 2016; Najem et al., 2026; Nelson, 2015), and is similarly employed in the present study to uncover distinct learner profiles. Our time-aware clustering identified five learner groups: Execution-Focused Digital Learners (EFDL), Low Engagement Digital Learners (LEDL), Reactive Strategic Digital Learners (RSDL), Consistent Digital Learners (CDL), and Independent Digital Learners (IDL). From an SRL perspective, CDL reflects more consistent and structured engagement pattern over time, IDL exhibits more deliberately planned engagement, while EFDL and RSDL show more execution-focused or reactive behaviours.

This is partially consistent with prior research, structured and plan-driven behavioural patterns are associated with more favourable outcomes (Bey & Champagnat, 2022), particularly when supported by early preparation, as observed in CDL learners. However, the present findings also nuance this view, showing that less structured or more reactive patterns, such as those observed in LEDL, RSDL, and IDL, can also lead to successful outcomes in a substantial number of cases. Accordingly, the relationship between behavioural profiles and performance is not strictly linear, as different engagement patterns can lead to both successful and at-risk outcomes. In particular, LEDL learners demonstrate that low digital activity does not necessarily imply poor performance, which aligns with participation-based clustering approaches (Nelson, 2015). This suggests that some learning strategies may be less visible in digital traces rather than less effective. From another SRL interpretation, the identified patterns can be linked to surface, deep, and achieving strategies (Fan et al., 2021). In the present results, LEDL behaviour aligns more closely with surface-level digital engagement, while CDL and partly IDL reflect deeper patterns, and EFDL, RSDL and partly IDL are more consistent with achieving-oriented strategies.

Beyond clustering, process models provide insight into the structure of learning behaviour, with transition patterns capturing pathways beyond aggregated metrics and enabling the examination of SRL through sequential dynamics. In the present results, EFDL, CDL, and IDL learners show comparable engagement with course-related activities; however, only CDL and IDL are associated with more favourable outcomes, while EFDL includes a higher proportion of at-risk students. This suggests that structurally similar behavioural patterns may correspond to different levels of regulation, consistent with prior finding (Maldonado-Mahauad et al., 2018), and highlights the need to consider behavioural structure alongside temporal and outcome-based dimensions. Accordingly, learning effectiveness is not determined solely by activity volume (Hernández-García et al., 2024), but by the organisation, timing, and context of actions. In addition, RSDL is characterised by transitions originating from task-related activities, whereas other clusters primarily converge towards them, indicating differences in how task engagement is integrated into the learning process, similar to temporal network approaches (Saqr et al., 2026).

Accordingly, dropout risk is best understood as a multi-phase process rather than a single critical event. Both overly simple and overly complex behavioural patterns are associated with less favourable outcomes (De Bruin et al., 2023), further highlighting the role of structural organisation. Consistent with prior work on behavioural continuity (Al-Hafdi & Alhalafawy, 2026), sustained engagement is generally associated with more favourable outcomes; however, its impact varies across phases, with early-stage activity acting as a potential protective factor. At the same time, no single learning strategy is universally ineffective, as most clusters include successful students, highlighting the need for nuanced interpretation. Moreover, the high performance of IDL students suggests that autonomy and flexibility may enhance success when supported by sufficient motivation (Xiao & Hew, 2024). The applied GAM model further supports these findings by capturing both structural and temporal dimensions of learning behaviour, indicating that underperformance reflects evolving behavioural patterns rather than isolated factors.

In response to RQ1, the analysis combining temporal segmentation and PM revealed distinct patterns of students' digital learning behaviour, characterised by differences in activity sequences, transition structures, and their evolution over time. These findings show that learning behaviour is shaped not only by activity levels, but by the organisation and temporal unfolding of actions. The identified learner profiles illustrate how these characteristics manifest in practice, including relatively consistent engagement in CDL, more independently organised patterns in IDL, execution-focused or reactive patterns in EFDL and RSDL, and lower levels of digital activity in LEDL. Overall, the results demonstrate that LMS event log data can systematically uncover temporal and process-based learning patterns beyond simple frequency measures.

In response to RQ2, the results show that these temporal and process-oriented patterns can be used to identify critical learning trajectories and time periods associated with academic risk. Differences across clusters reveal that students follow distinct learning strategies over the semester, which are reflected in both their behavioural sequences and their temporal distribution of activity. The findings indicate that dropout risk is not tied to a single behaviour, but emerges from the interaction of engagement patterns across different course periods. In particular, early and sustained engagement appears to have a protective role, while irregular or delayed activity is more often associated with risk. Process-based indicators, such as model simplicity, further capture how the organisation of learning behaviour contributes to outcomes. Overall, the results suggest that underperformance is best understood as a dynamic process shaped by evolving learning trajectories and time-dependent behavioural differences rather than isolated events.

Theoretical contributions

This study contributes to the understanding of SRL by showing that it can be interpreted as a temporally evolving, process-oriented phenomenon reflected in behavioural traces. The findings suggest that SRL cannot be captured by a single behavioural dimension, as similar observable behaviours may reflect distinct levels of regulation; however, incorporating multiple perspectives allows these patterns to be differentiated. In addition, the results indicate that SRL varies across phases of the learning process, with early and sustained engagement playing a key role, highlighting the importance of temporal and process-based approaches.

Practical implications

From a practical perspective, the study provides an analytical framework that integrates traditional data analysis, temporal segmentation, and process mining techniques to better understand student learning behaviour. This approach enables a more comprehensive interpretation of engagement patterns and may support early-warning systems, as different phases of the semester indicate evolving dropout risk. At the same time, implementation requires careful consideration of institutional and ethical aspects. The results may support students' self-evaluation in line with SRL and provide instructors with actionable insights into intervention points. In addition, the findings may inform curriculum design by highlighting critical phases and behavioural patterns.

Limitation and future directions

This study has several limitations, since it is based on a single course at one university, which limits generalisability. The relatively small sample size (n = 103), particularly the low number of at-risk students, may affect the robustness of the predictive results, as predictive models are sensitive to low dropout rates (Rebelo Marcolino et al., 2025). In addition, the analysis relies on Moodle log data, which capture only digital traces and do not account for offline learning, peer interaction, or motivational factors. As the study is based on observational data from a single semester, the relationship between behaviour and outcomes cannot be interpreted causally, and unobserved factors such as prior knowledge may influence the results.

Future research should validate these findings across multiple courses and contexts using larger samples. Given the sensitivity of predictive approaches to low dropout rates, further testing on larger datasets is particularly important. Integrating additional data sources, such as surveys or qualitative measures, could provide deeper insight into the relationship between behavioural patterns and learning processes. Another promising direction is the development of real-time predictive approaches to identify at-risk trajectories and support timely interventions.

Conclusion

This study shows that combining temporal analysis, clustering, and process-oriented modelling provides a more nuanced understanding of digital learning behaviour and its relation to SRL. The findings indicate that learning is shaped not only by activity levels, but by the timing, structure, and evolution of engagement, with different behavioural patterns reflecting distinct forms of regulation. Early and sustained engagement, as well as well-structured processes, appear to play a key role, while lower digital activity does not necessarily indicate weaker SRL. Methodologically, the study demonstrates the value of integrating temporal and process-based approaches to interpret behavioural traces and identify risk patterns. From a practical perspective, these insights may support more proactive, data-informed student support and the development of adaptive learning environments.

Appendix

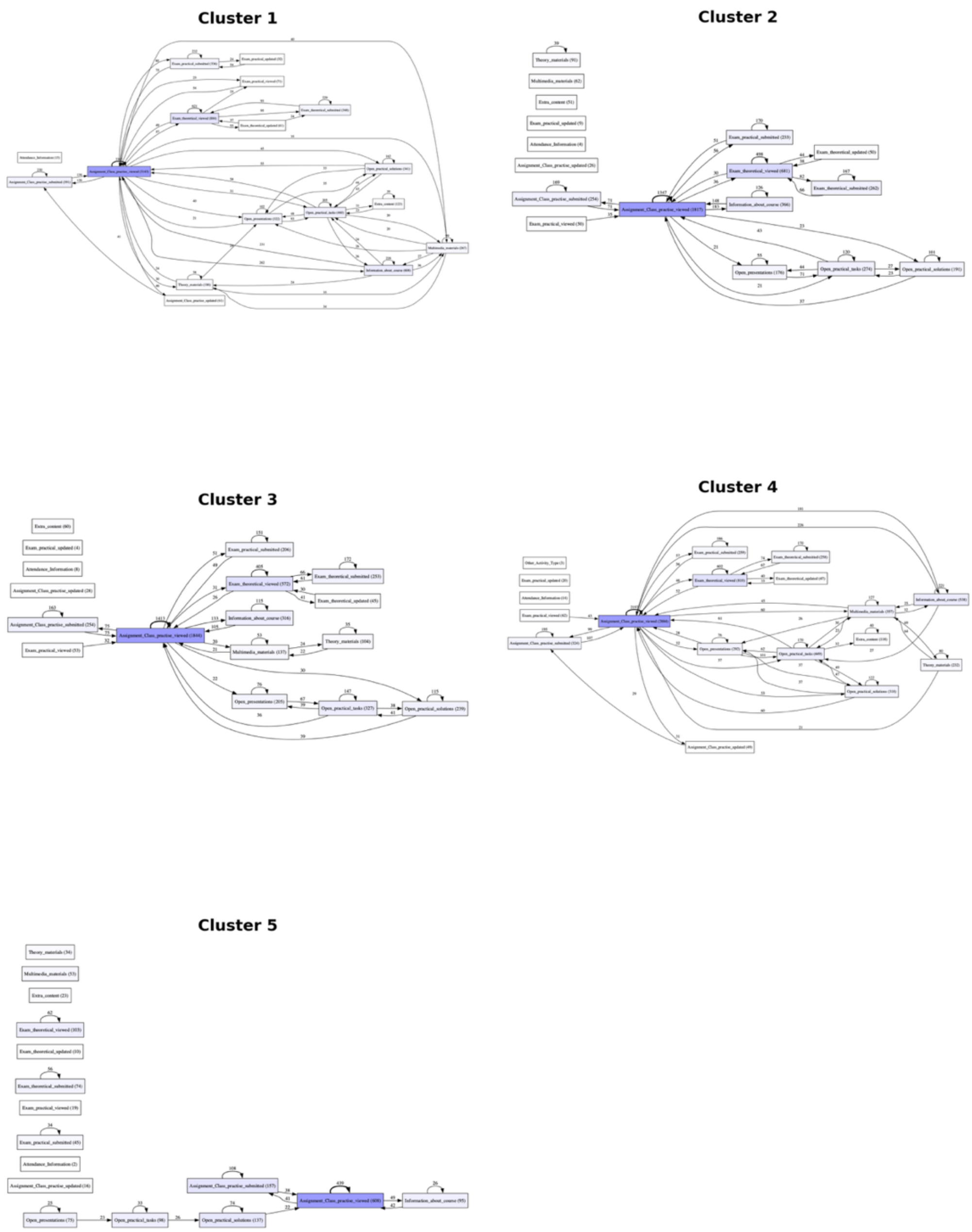
Figure 9

Activity categories, subcategories and their definitions (prepared by the authors)

Activity category	Activity subcategory	Definition
Assignment activities	Assignment class practise submitted/updated/viewed	Individual assignment tasks, including access, submission, and updates.
Attendance	Attendance information	Viewing the attendance information.
Exam activities	Exam practical submitted/updated/viewed Exam theoretical submitted/updated/viewed	Exam-related activities, including access, submission, and updates of both practical and theoretical tasks.
Extra content	Extra content	Supplementary materials for those interested. Viewing is optional.
General information	Information about course	Activities that connect to viewing general information about the course and the requirements.
Multimedia content	Multimedia materials	Activities connected to video content about the practical and theoretical classes.
Presentation	Open presentations Theory materials	Activities related to written materials for both practical tasks and theoretical explanations.
Tasks	Open practical solutions Open practical tasks Other activity type	Activities related to practical programming tasks, including solutions and other unspecified task types.

Figure 10

Frequency-based DFGs for each cluster after applying a minimum frequency threshold (prepared by the authors)



Abbreviations

Learning Management Systems: LMS; Process Mining: PM; Educational process mining: EPM; Self-regulated learning: SRL; Educational data mining: EDM; Learning analytics: LA; Generalised Additive Model: GAM; Local Interpretable Model-Agnostic Explanations: LIME; Dynamic time warping: DTW; Calinski-Harabasz index: CH index; Davies–Bouldin Index: DB index; Directly-follow graph: DFG; Alpha Miner algorithm: AM; Inductive Miner algorithm: IM; Heuristic Miner algorithm: HM; Effective Degrees of Freedom: Effective DoF; Corrected Akaike Information Criterion: AICc; Variance Inflation Factor: VIF; Execution-Focused Digital Learners: EFDL; Low Engagement Digital Learners: LEDL; Reactive Strategic Digital Learners: RSDL; Consistent Digital Learners: CDL; Independent Digital Learners: IDL

Author's contributions

BSO and SF contributed equally to this work. Both authors jointly conceived the study, designed the research methodology, interpreted the findings, and wrote and revised the manuscript. BSO was primarily responsible for data collection, data preprocessing, software implementation, and conducting the empirical analyses. SF primarily contributed to the conceptual development of the study, the theoretical framework, methodological refinement, and supervision of the research. Both authors critically reviewed the manuscript and approved the final version.

Author's information

BSO is a PhD candidate at the Department of Computer Science, Corvinus University of Budapest. Her research focuses on educational data mining, process mining, learning analytics, and artificial intelligence in higher education. SF is Full Professor and Head of the Department of Computer Science at Corvinus University of Budapest. Her research interests include educational data mining, learning analytics, artificial intelligence, business analytics, and process mining.

Funding

The project identified by EKOP-CORVINUS-24-3-017 was realised with the support of the National Research, Development, and Innovation Fund provided by the Ministry of Culture and Innovation, as part of the University Research Scholarship Program announced for the 2024/2025 academic year.

Availability of data and materials

The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request, subject to institutional and ethical considerations regarding student data.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

All authors are affiliated with the Department of Computer Science, Corvinus University of Budapest, Hungary.
Szabina Fodor: Full professor, szabina.fodor@uni-corvinus.hu, https://orcid.org/0000-0001-5459-7912
Bernadett Sarró-Oláh: PhD candidate, bernadett.sarro-olah@uni-corvinus.hu, https://orcid.org/0009-0009-2940-3376

Received: 27 August 2025 Accepted: 13 July 2026
Published online: 1 January 2027 (Online First: 8 September 2026)

References

Aalst, W. M. P. van der, Dongen, B. F. van, Günther, C. W., Rozinat, A., Verbeek, H. M. W., & Weijters, A. J. M. M. (2009). ProM: The process mining toolkit. In A. K. Alves De Medeiros & B. Weber (Eds), *Proceedings of the BPM 2009 Demonstration Track (BPMDemos 2009, Ulm, Germany, September 8, 2009)* (Vol. 489, pp. 1–4). CEUR-WS.org. https://ceur-ws.org/Vol-489/paper3.pdf

Adigun, O. T., Mpofu, N., & Maphalala, M. C. (2025). Fostering self-directed learning in blended learning environments: A constructivist perspective in Higher Education. *Higher Education Quarterly*, *79*(1), e12572. https://doi.org/10.1111/hequ.12572

Ajjawi, R., Dracup, M., Zacharias, N., Bennett, S., & Boud, D. (2020). Persisting students' explanations of and emotional responses to academic failure. *Higher Education Research & Development*, *39*(2), 185–199. https://doi.org/10.1080/07294360.2019.1664999

Al-Hafdi, F. S., & Alhalafawy, W. S. (2026). Learning analytics for reducing student dropout in digital video platforms. *PeerJ Computer Science*, *12*, e3532. https://doi.org/10.7717/peerj-cs.3532

AlQaheri, H., & Panda, M. (2022). An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality. *Information*, *13*(1), 29. https://doi.org/10.3390/info13010029

Alseitova, A., Oliveira, W., Li, Z., & Hamari, J. (2025). Understanding Students' Behavior in Learning Management Systems Through Their Personality Traits. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-025-09931-w

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2022). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, *55*(6), 3026–3054. https://doi.org/10.3758/s13428-022-01939-9

Baker, R., & Siemens, G. (2014). Educational Data Mining and Learning Analytics. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd edn, pp. 253–272). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016

Berti, A., Van Zelst, S., & Schuster, D. (2023). PM4Py: A process mining library for Python. *Software Impacts*, *17*, 100556. https://doi.org/10.1016/j.simpa.2023.100556

Bessadok, A., Abouzinadah, E., & Rabie, O. (2023). Exploring students digital activities and performances through their activities logged in learning management system using educational data mining approach. *Interactive Technology and Smart Education*, *20*(1), 58–72. https://doi.org/10.1108/ITSE-08-2021-0148

Bey, A., & Champagnat, R. (2022). Analyzing Student Programming Paths using Clustering and Process Mining: *Proceedings of the 14th International Conference on Computer Supported Education*, 76–84. https://doi.org/10.5220/0011077300003182

Bhutoria, A. (2022). Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 11–15. https://doi.org/10.1145/2567574.2567604

Bondaug, F. G., & Tubo, B. F. (2024). High-dimensional sparse predictive modeling applied to varied correlation structure via the generalized additive model. *Procedia Computer Science*, *234*, 70–78. https://doi.org/10.1016/j.procs.2024.02.153

Brochenin, R., Buijs, J., Vahdat, M., & van der Aalst, W. (2017). *Resource Usage Analysis from a Different Perspective on MOOC Dropout* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1710.05917

Buijs, J. C. A. M., Van Dongen, B. F., & Van Der Aalst, W. M. P. (2014). Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity. *International Journal of Cooperative Information Systems*, *23*(01), 1440001. https://doi.org/10.1142/S0218843014400012

Cenka, B. A. N., Santoso, H. B., & Junus, K. (2022). Analysing student behaviour in a learning management system using a process mining approach. *Knowledge Management & E-Learning: An International Journal*, 62–80. https://doi.org/10.34105/j.kmel.2022.14.005

Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, *32*(1), 74–88. https://doi.org/10.1007/s12528-019-09225-y

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42–54. https://doi.org/10.1016/j.compedu.2016.02.006

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., & Caruana, R. (2021). How Interpretable and Trustworthy are GAMs? *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 95–105. https://doi.org/10.1145/3447548.3467453

Chanifah, S., Andreswari, R., & Fauzi, R. (2021). Analysis of Student Learning Pattern in Learning Management System (LMS) using Heuristic Mining a Process Mining Approach. *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, 121–125. https://doi.org/10.1109/ICERA53111.2021.9538654

Chen, F., & Cui, Y. (2020). Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *Journal of Learning Analytics*, *7*(2), 1–17. https://doi.org/10.18608/jla.2020.72.1

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two Decades of Artificial Intelligence in Education. *Educational Technology & Society*, *25*(1), 28–47. JSTOR.

Cristea, T. S., Heikkinen, S., Snijders, C., Saqr, M., Matzat, U., Conijn, R., & Kleingeld, A. (2025). Dynamics of self-regulated learning: The effectiveness of students' strategies across course periods. *Computers & Education*, *228*, 105233. https://doi.org/10.1016/j.compedu.2025.105233

De Bruin, A. B. H., Biwer, F., Hui, L., Onan, E., David, L., & Wiradhany, W. (2023). Worth the Effort: The Start and Stick to Desirable Difficulties (S2D2) Framework. *Educational Psychology Review*, *35*(2), 41. https://doi.org/10.1007/s10648-023-09766-w

Detmer, A. R., Ward, E. J., Hunsicker, M. E., Andrews, K. S., Conrad, M., Ferriss, B. E., Hazen, E. L., Holsman, K. K., Indivero, J., Large, S. I., Malick, M., Marshall, K. N., Munsch, S. H., Oken, K. L., Satterthwaite, W. H., Shotwell, S. K., Thompson, A. R., & Samhouri, J. F. (2025). Evaluating the robustness of generalized additive models as a tool for threshold detection in variable environments. *Ecosphere*, *16*(3), e70117. https://doi.org/10.1002/ecs2.70117

Fan, Y., Matcha, W., Uzir, N. A., Wang, Q., & Gašević, D. (2021). Learning Analytics to Reveal Links Between Learning Design and Self-Regulated Learning. *International Journal of Artificial Intelligence in Education*, *31*(4), 980–1021. https://doi.org/10.1007/s40593-021-00249-z

Feng, G., Fan, M., & Ao, C. (2022). Exploration and Visualization of Learning Behavior Patterns From the Perspective of Educational Process Mining. *IEEE Access*, *10*, 65271–65283. https://doi.org/10.1109/ACCESS.2022.3184111

Garrison, D. R. (1997). Self-Directed Learning: Toward a Comprehensive Model. *Adult Education Quarterly*, *48*(1), 18–33. https://doi.org/10.1177/074171369704800103

Gasevic, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. *Journal of Learning Analytics*, *4*(2). https://doi.org/10.18608/jla.2017.42.10

Ghanem, K. (2013). Towards More Accurate Clutering Method by Using Dynamic Time Warping. *International Journal of Data Mining & Knowledge Management Process*, *3*(2), 107–118. https://doi.org/10.5121/ijdkp.2013.3207

Hachicha, W., Ghorbel, L., Champagnat, R., Zayani, C. A., & Amous, I. (2021). Using Process Mining for Learning Resource Recommendation: A Moodle Case Study. *Procedia Computer Science*, *192*, 853–862. https://doi.org/10.1016/j.procs.2021.08.088

Hartmann, C., Rummel, N., & Bannert, M. (2022). Using HeuristicsMiner to Analyze Problem-Solving Processes: Exemplary Use Case of a ProductiveFailure Study. *Journal of Learning Analytics*, *9*(2), 66–86. Scopus. https://doi.org/10.18608/jla.2022.7363

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3). https://doi.org/10.1214/ss/1177013604

Hastie, T., & Tibshirani, R. (2014). Generalized Additive Models. In R. S. Kenett, N. T. Longford, W. W. Piegorsch, & F. Ruggeri (Eds), *Wiley StatsRef: Statistics Reference Online* (1st edn). Wiley. https://doi.org/10.1002/9781118445112.stat03141

Helal, D., Helmy, Y., & Elzanfaly, D. (2025). Predicting student academic performance using engagement features: A process mining and deep learning approach. *Journal of Theoretical and Applied Information Technology*, *103*(9), 3803–3820.

Hemmert, G. A. J., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2018). Log-likelihood-based Pseudo- R^2 in Logistic Regression: Deriving Sample-sensitive Benchmarks. *Sociological Methods & Research*, *47*(3), 507–531. https://doi.org/10.1177/0049124116638107

Hernández-García, Á., Cuenca-Enrique, C., Del-Río-Carazo, L., & Iglesias-Pradas, S. (2024). Exploring the relationship between LMS interactions and academic performance: A Learning Cycle approach. *Computers in Human Behavior*, *155*, 108183. https://doi.org/10.1016/j.chb.2024.108183

Hinds, J., & Joinson, A. (2019). Human and Computer Personality Prediction From Digital Footprints. *Current Directions in Psychological Science*, *28*(2), 204–211. https://doi.org/10.1177/0963721419827849

Jin, S.-H., Im, K., Yoo, M., Roll, I., & Seo, K. (2023). Supporting students’ self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education*, *20*(1), 37. https://doi.org/10.1186/s41239-023-00406-5

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. https://doi.org/10.1016/j.chb.2017.12.015

Kalita, E., Oyelere, S. S., Gaftandzhieva, S., Rajesh, K. N. V. P. S., Jagatheesaperumal, S. K., Mohamed, A., Elbarawy, Y. M., Desuky, A. S., Hussain, S., Cifci, M. A., Theodorou, P., Hilčenko, S., Hazarika, J., & Ali, T. (2025). Educational data mining: A 10-year review. *Discover Computing*, *28*(1), 81. https://doi.org/10.1007/s10791-025-09589-z

Koschke, R., & Steinbeck, M. (2020). Clustering Paths With Dynamic Time Warping. *2020 Working Conference on Software Visualization (VISSOFT)*, 89–99. https://doi.org/10.1109/VISSOFT51673.2020.00014

Lassen, K. B., & Van Der Aalst, W. M. P. (2009). Complexity metrics for Workflow nets. *Information and Software Technology*, *51*(3), 610–626. https://doi.org/10.1016/j.infsof.2008.08.005

Leemans, S. J. J., Fahland, D., & Van Der Aalst, W. M. P. (2014). Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In N. Lohmann, M. Song, & P. Wohed (Eds), *Business Process Management Workshops* (Vol. 171, pp. 66–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-06257-0_6

Leemans, S. J. J., Fahland, D., & Van Der Aalst, W. M. P. (2018). Scalable process discovery and conformance checking. *Software & Systems Modeling*, *17*(2), 599–631. https://doi.org/10.1007/s10270-016-0545-x

Li, Y., Jiang, Q., Xiong, W., & Zhao, W. (2023). Investigating behavior patterns of students during online self-directed learning through process mining. *Education and Information Technologies*, *28*(12), 15765–15787. https://doi.org/10.1007/s10639-023-11830-5

Lin, H.-M., Wu, J.-Y., Liang, J.-C., Lee, Y.-H., Huang, P.-C., Kwok, O.-M., & Tsai, C.-C. (2023). A review of using multilevel modeling in e-learning research. *Computers & Education*, *198*, 104762. https://doi.org/10.1016/j.compedu.2023.104762

López-Rubio, E., Palomo, E. J., & Ortega-Zamorano, F. (2018). Unsupervised learning by cluster quality optimization. *Information Sciences*, *436–437*, 31–55. https://doi.org/10.1016/j.ins.2018.01.007

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior*, *80*, 179–196. https://doi.org/10.1016/j.chb.2017.11.011

Martínez-Carrascal, J. A., Munoz-Gama, J., & Sancho-Vinuesa, T. (2024). Evaluation of Recommended Learning Paths Using Process Mining and Log Skeletons: Conceptualization and Insight into an Online Mathematics Course. *IEEE Transactions on Learning Technologies*, *17*, 555–568. https://doi.org/10.1109/TLT.2023.3298035

Mukala, P. (2025). *Leveraging Process Mining, Learning Analytics, and AI to Explore Learning Behaviors in MOOCs and FLOSS Communities: A Comprehensive Mise-au-Point of Key Findings*. Preprints. https://doi.org/10.36227/techrxiv.173835393.35478241/v1

Najem, K., Seghroucheni, Y. Z., IPSS, Faculty of Science, Mohammed V university in Rabat, Rabat, Morocco, Ziti, S., & IPSS, Faculty of Science, Mohammed V university in Rabat, Rabat, Morocco. (2026). Behavioral Clustering for Adaptive Learning: A Data-Driven Alternative to Static Learning Style Models. *International Journal of Information and Education Technology*, *16*(1), 196–204. https://doi.org/10.18178/ijiet.2026.16.1.2494

Nelson, K. (2015). USING K-MEANS CLUSTERING TO MODEL STUDENTS LMS PARTICIPATION IN TRADITIONAL COURSES. *Issues In Information Systems*. https://doi.org/10.48009/4_iis_2015_102-110

Ngah, A. H., Kamalrulzaman, N. I., Mohamad, M. F. H., Rashid, R. A., Harun, N. O., Ariffin, N. A., & Osman, N. A. A. (2022). The sequential mediation model of students' willingness to continue online learning during the COVID-19 pandemic. *Research and Practice in Technology Enhanced Learning*, *17*(1), 13. https://doi.org/10.1186/s41039-022-00188-w

Ngo, Q.-D., & Phan, L.-H. (2025). Analysis of multivariate indoor building data: A comparative study of time-series clustering methods. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2025.3596297

Nizam Ismail, S., Hamid, S., Ahmad, M., Alaboudi, A., & Jhanjhi, N. (2021). Exploring Students Engagement Towards the Learning Management System (LMS) Using Learning Analytics. *Computer Systems Science and Engineering*, *37*(1), 73–87. https://doi.org/10.32604/csse.2021.015261

Ortiz-Vilchis, P., & Ramirez-Arellano, A. (2023). Learning Pathways and Students Performance: A Dynamic Complex System. *Entropy*, *25*(2), 291. https://doi.org/10.3390/e25020291

Palanci, A., Yılmaz, R. M., & Turan, Z. (2024). Learning analytics in distance education: A systematic review study. *Education and Information Technologies*, *29*(17), 22629–22650. https://doi.org/10.1007/s10639-024-12737-5

Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, *17*, 49–64.

Patel, N., Sellman, C., & Lomas, D. (2017). *Mining Frequent Learning Pathways from a Large Educational Dataset* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1705.11125

Pentland, B. T., Recker, J., Wolf, J., & Wyner, G. (2020). Bringing Context Inside Process Research with Digital Trace Data. *Journal of the Association for Information Systems*, *21*(5), 1214–1236. https://doi.org/10.17705/1jais.00635

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. https://doi.org/10.1016/j.patcog.2010.09.013

Real, E. M., Pinheiro Pimentel, E., De Oliveira, L. V., Cristina Braga, J., & Stiubiener, I. (2020). Educational Process Mining for Verifying Student Learning Paths in an Introductory Programming Course. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE44824.2020.9274125

Real, E., & Pimentel, E. (2025). An Educational Process Mining Model on Students' Paths Data from Virtual Learning Environments. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-025-09824-y

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. *Scientific Reports*, *15*(1), 9840. https://doi.org/10.1038/s41598-025-93918-1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

Saha, M., Islam, S., Akhi, A. A., & Saha, G. (2022). *Factors Affecting Success and Failure in Higher Education Mathematics: Students and Teachers' Perspectives*. https://doi.org/10.20944/preprints202209.0378.v1

Saint, J., Whitelock-Wainwright, A., Gasevic, D., & Pardo, A. (2020). Trace-SRL: A Framework for Analysis of Microlevel Processes of Self-Regulated Learning From Trace Data. *IEEE Transactions on Learning Technologies*, *13*(4), 861–877. https://doi.org/10.1109/TLT.2020.3027496

Saqr, M., López-Pernas, S., & Törmänen, T. (2026). A temporal network approach to reveal the longitudinal dynamics of CSCL group regulation and productive collaboration. *International Journal of Computer-Supported Collaborative Learning*. https://doi.org/10.1007/s11412-025-09464-5

Saqr, M., López-Pernas, S., Törmänen, T., Kaliisa, R., Misiejuk, K., & Tikka, S. (2025). Transition Network Analysis: A Novel Framework for Modeling, Visualizing, and Identifying the Temporal Patterns of Learners and Learning Processes. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 351–361. https://doi.org/10.1145/3706468.3706513

Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter*, *25*(1), 36–42. https://doi.org/10.1145/3606274.3606278

Semler, R. F., Semler, J. R., Wehrmeister, M. A., Southier, L. F. P., Casanova, D., & Teixeira, M. (2025). Educational process mining: Literature classification, gaps, and emerging opportunities. *Frontiers in Education*, *10*, 1543761. https://doi.org/10.3389/feduc.2025.1543761

Servén, D., Brummitt, C., Király, F., JodesL, Ouslan, A., Miccoli, F., IlyesD, Egan, J., Lapenta, L., Servén Marín, L., Ranjan, S., Rose, S. A., Fasci, U., Szépe, V., hlink, & Abedi, H. (2025). *Dswah/pyGAM: V0.10.1* (Version v0.10.1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1208723

Shirk, J. (2020). Designing a Self-Paced Learning Experience to Support Learner Self-Regulation. *Journal of Teaching and Learning with Technology*, *9*(1). https://doi.org/10.14434/jotlt.v9i1.29527

Sridharan, T. B., & Sharagaharajan Akilashri, D. P. S. (2025). Multimodal learning analytics for students behavior prediction using multi-scale dilated deep temporal convolution network with improved chameleon Swarm algorithm. *Expert Systems with Applications*, *286*, 128113. https://doi.org/10.1016/j.eswa.2025.128113

Susnjak, T., Ramaswami, G. S., & Mathrani, A. (2022). Learning analytics dashboard: A tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, *19*(1), 12. https://doi.org/10.1186/s41239-021-00313-7

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., & Woods, E. (2020). Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, *21*(118), 1–6.

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning*, *10*(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022

Van Der Aalst, W. (2016). *Process Mining*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

Van Der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47

Vázquez-Barreiros, B., Mucientes, M., & Lama, M. (2015). ProDiGen: Mining complete, precise and minimal structure process models with a genetic algorithm. *Innovative Applications of Artificial Neural Networks in Engineering*, *294*, 315–333. https://doi.org/10.1016/j.ins.2014.09.057

Wang, S. (2025). Analysis of Learning Behaviour of Open Education Learners based on Principal Component Analysis and K-means Clustering Algorithm. *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, 1–7. https://doi.org/10.1109/ICISCN64258.2025.10934291

Weijters, A., Aalst, W., & Medeiros, A. (2006). Process Mining with the Heuristics Miner-algorithm. In *Cirp Annals-manufacturing Technology—CIRP ANN-MANUF TECHNOL* (Vol. 166).

Wijnhoven, F., Hoffmann, P., Bemthuis, R., & Boksebeld, J. (2023). Using process mining for workarounds analysis in context: Learning from a small and medium-sized company case. *International Journal of Information Management Data Insights*, *3*(1), 100163. https://doi.org/10.1016/j.jjimei.2023.100163

Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, *1*.

Wood, S. N. (2025). Generalized Additive Models. *Annual Review of Statistics and Its Application*, *12*(1), 497–526. https://doi.org/10.1146/annurev-statistics-112723-034249

Xiao, Y., & Hew, K. F. (2024). Personalized gamification versus one-size-fits-all gamification in fully online learning: Effects on student motivational, behavioral and cognitive outcomes. *Learning and Individual Differences*, *113*, 102470. https://doi.org/10.1016/j.lindif.2024.102470

Zaharuddin, Chen Yu, & Yao, G. (2024). Enhancing Student Engagement with AI-Driven Personalized Learning Systems. *International Transactions on Education Technology (ITEE)*, *3*(1), 1–8. https://doi.org/10.33050/itee.v3i1.662

Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., & Caruana, R. (2018). *Axiomatic Interpretability for Multiclass Additive Models* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1810.09092

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Publisher's Note