Who is My Audience?
Discovering Student Personas from LMS Data

Ahmad Ainun Herlambang[a]; Abel Armas-Cervantes[a]; Antonette Mendoza[a]; Nungki Dian S Darmayanti[b]
The University of Melbourne[a], National Research and Innovation Agency of Indonesia[b]
Corresponding Author Email: abel.armas@unimelb.edu.au

ABSTRACT

CONTEXT

Academics often struggle to grasp a complete picture of student cohorts. This can be due to several reasons, such as size of the cohorts, little interaction with students, and lecture absenteeism. Having a complete picture of a cohort can help academics detect desired and abnormal behaviours, which can inform better teaching and learning practices. Learning Management Systems (LMS) are commonly used platforms where academics upload content, create exercises, and communicate with students. Hence, LMS capture a vast amount of student interactions and can be leveraged to help academics understand the students' learning behaviours.

PURPOSE OR GOAL

This paper presents a pipeline to analyse the data captured in LMS for the creation of personas. A persona represents a group of students with similar learning behaviours. The goal is to use the presented pipeline to extract a handful of personas that are representatives of a large cohort, and that can provide academics with key information about the cohort (e.g., average number of study hours, compliance with the teaching strategy and other study behaviours). The personas are extracted solely from the data collected in the LMS; hence, they can be derived easily at various points during the study period.

APPROACH OR METHODOLOGY/METHODS

The raw LMS data was pre-processed to filter out dropouts, the final data contained the interactions of 185 students. From the data, 12 features were defined capturing learning tactics, content recap, compliance with flipped classroom strategy, and preferred learning hours. These features were used as input for a clustering algorithm (K-Means and DBSCAN), which group students with similar behavioural profiles according to the 12 features defined. Each cluster was summarised into a persona, which was enriched with charts showing statistics of the clustered group for interpretation and comparison purpose.

ACTUAL OR ANTICIPATED OUTCOMES

The research presented a reusable pipeline to analyse LMS data. The pipeline presents a generic set of steps that can be reused and adapted to analyse data from any LMS. In the presented study, we used our pipeline to analyse a semester's worth of data and discovered five personas, each reflecting distinct differences in student activity styles and learning behaviours.

CONCLUSIONS/RECOMMENDATIONS/SUMMARY

The paper demonstrates that LMS data can be exploited to gain valuable insights into student behaviours using personas. These behaviours can be extracted by defining meaningful features that captures student learning patterns. Although the pipeline is reusable, the optimal number of personas may vary depending on the nature of the dataset. The personas can give academics a summary of the behaviour of the student cohort and inform their teaching improvement initiatives.

KEYWORDS

Process mining, Machine learning, LMS data, Persona

Introduction

Academics often struggle to understand how students interact with learning materials. This problem is heightened in cohorts with hundreds of students where academics do not have an overview of how the cohort is continuously engaging with the content. This lack of visibility makes it difficult to meet student needs, especially when attendance and engagement are low. This is critical as student engagement is strongly linked to academic performance (Matcha et al., 2020).

Post-COVID-19, student activity has shifted online, with Learning Management Systems (LMS) becoming the primary platform for engagement. Although LMS capture extensive interaction data, built-in dashboards often provide only surface-level analytics (Susnjak et al., 2022), leaving patterns like content recap or study time preferences unexplored. These deeper behavioural patterns can help us understand how students engage with learning materials, which can inform academics to provide personalised feedback, timely interventions, and improved course design.

This paper investigates how data recorded in LMS can be leveraged to gain insights into student behavioural patterns to support teaching and learning. We aim to answer the following research question: How can LMS data be leveraged to understand student engagement patterns?

To address the question above, we use process mining (PM) (van der Aalst, 2012) and machine learning (ML) techniques to analyse the LMS data. PM is a family of tools and techniques that analyse data through a process lens. In this paper, PM is used to reconstruct behavioural flows captured in LMS event logs, revealing learning pathways and engagement patterns (Armas-Cervantes et al., 2023; Song et al., 2024). ML, on the other hand, offers a range of tools that can be used for data analysis (Guyon et al., 2019). In the context of this paper, ML techniques for dimensionality reduction and clustering are used to group together students with similar learning patterns. These patterns are defined as *features* extracted from the LMS data.

To showcase our approach, we use a real-life dataset containing a semester's worth of data. This dataset contains the interactions of 185 students with Canvas LMS at the University of Melbourne. The use of the dataset was approved under Human Ethics Student Coursework Exemption, as part of the project "Master Research Project - LMS Analytics" (Project ID: 32636).

The structure of the paper is as follows. The following sections presents the literature review and background, then it is followed by the methodology, experiments and results. We follow with a discussion about the results before we conclude the paper.

Literature Review and Background

Learning Patterns vs Academic Performances

Several studies have focused on the impact of learning strategies and behaviours on academic outcomes. Diverse learning tactics (e.g., reading, exercises, videos) are linked to better performance (Matcha et al., 2020), with intensive strategies showing the most positive effects, though not always significant (Song et al., 2024). While longer study durations often correlate with achievement, excessive study (over 5–10 hours) may lead to diminishing returns (Liu, 2022), suggesting the importance of when and how students' study. For instance, morning learners (early chronotypes) tend to perform better (Enright & Refinetti, 2017). Repetition and reengagement suggest deeper learning and memory retention (Musfeld et al., 2023), while pre- and post-lecture activities aid preparation and consolidation in flipped classrooms. These environments have been shown to improve performance and motivation (Gonda et al., 2021). However, most studies rely on survey-based or experimental data. Few studies integrate these behavioural factors into scalable, data-driven approaches. The present study addresses this gap by extracting pedagogically relevant features from LMS logs to generate interpretable behavioural student profiles.

Personas

Personas are fictional characters that represent archetypical users. These personas are used as tools to represent groups of people with similar characteristics and behaviours. Traditionally, personas are built from interviews or surveys as a one-time document, but they can become static and fail to reflect changes that occur over time, hence they can become outdated or misrepresent evolving people's needs (Brown, 2023). Recent works explores a data-driven approach in segmenting users through clustering, offering rich and dynamic insights for targeted decision-making (Park et al., 2024). The present paper uses a similar data-driven approach and analyses data collected in LMS to generate personas.

Clustering Algorithms

Clustering is a well-established and commonly used method in machine learning. It is an unsupervised method to group unlabelled data into subsets (clusters) based on similarity. Its main goal is to uncover hidden structures and naturally group similar instances (Jain, 2010). In this study, clustering serves as a tool to identify behavioural patterns in LMS data, LMS data into interpretable personas that reflect diverse student engagement styles.

Among various clustering techniques, this research focuses on K-Means and DBSCAN. K-Means is a widely used method due to its simplicity and efficiency. It assigns data points to a predefined number of clusters by minimising the sum of squared distances to centroids (Macqueen, 1967). Density-Based Spatial Clustering of Applications with Noise (DBSCAN), on the other hand, detects clusters based on density and does not require a predetermined number of clusters. It is also used to identify outliers and irregularly shaped clusters (Ester et al., 1996). Through the clustering techniques on the LMS data, this study produces personas that reflect student activity patterns in a scalable and interpretable format.

Methodology, Experiments and Results

Methodology

In this work we use a combination of techniques for the analysis of the LMS data: Process Mining (PM) (van der Aalst, 2012) and Machine Learning (ML). PM offers a range of tools and techniques to analyse data through a process lens and has been used in previous studies to discover student learning pathways captured in LMS data (Armas-Cervantes et al., 2023, 2024; Song et al., 2024). ML offers a range of methods for data analysis. Relevant to this paper, ML offers techniques for reducing dimensionality and identifying patterns in noisy data (Guyon et al., 2019). ML techniques have been used in the analysis of student data to predict academic outcomes (Agha et al. (2023).

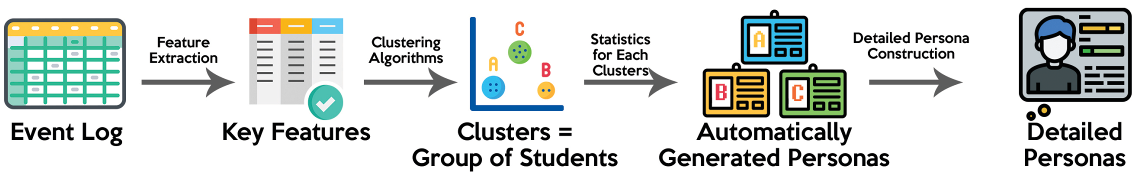

Figure 1. Research pipeline from data pre-processing to persona generation

Our proposed data analysis pipeline is presented in Figure 1. The input is an event log consisting of transactional records that capture students' interactions with the LMS. Each entry in the event log records an LMS resource visit. The entry contains information about the visited resource (e.g., quiz, video, reading material), a timestamp of when the resource was visited, and a case ID (unique student ID) relating all resources visited by the same student. Intuitively, these event logs describe the sequence of resources visited by an student throughout the semester.

Next, a range of features are extracted from the event log. These features summarise key aspects of student activities and were defined to capture patterns that are pedagogically relevant to academic performance. The selected features revolve around learning tactics and activity levels, covering interaction frequency, timing patterns (e.g., daytime vs nighttime study), types of learning modules accessed (e.g., lecture capture, content summary, quiz), and content repetition. To define some of the features, PM was used to analyse the order in which the LMS resources were visited. Having a process view of a student's learning pathway through the semester, it is possible to analyse when a particular set of resources were executed in the desired order, which is critical in some instructional approaches, such as flipped classroom where students are expected to review material prior to the flipped session.

The extracted features were then used as input to machine learning clustering algorithms (K-Means and DBSCAN) to segment the student cohort into clusters to represent students with similar features. Given that the number of features can be large, dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be used to aid the clustering techniques. After clustering, each resulting cluster was summarised into a data-driven persona, representing a behavioural archetype that captures the typical activity style within the group. Since personas are commonly used in business contexts to support user-centred design, this study adjusted the persona template to fit the educational context. The persona is enriched with cluster statistics such as average number of interactions by the cluster of students, module preferences, and temporal study habits.

Finally, a comprehensive descriptive analysis was conducted to interpret the statistical values associated with each automatically generated persona. This involved examining metrics such as average interaction counts, module access preferences, study time distributions, and content revisit frequencies within each cluster. These quantitative summaries were then used to develop rich, interpretable descriptions of each persona. Charts and diagrams were utilised to support comparisons between groups and to facilitate easier understanding by educators.

Dataset description

We use the pipeline presented in Figure 1 to analyse a real-life dataset from the LMS. The dataset captures student interactions during a whole semester for a subject that is part of the Master's course at the University of Melbourne. The original dataset consisted of interactions from 194 students; however, 9 students were filtered out because they dropout from the subject before the third week, then only 185 students were left in the final dataset. The log captures 16 weeks of activity, 12 weeks of teaching and 4 weeks of exam preparation. The subject was delivered in a flipped classroom format, where the students were expected to watch videos, read modules, attend seminars, complete workshops, and take weekly quizzes within the LMS. The dataset included key attributes such as session ID, academic week, timestamps, modules accessed, and resource type. Synthetic entries labelled "Live Zoom Seminar" were added to mark lecture dates not captured by the LMS, as live Zoom sessions were conducted outside the system's logging capabilities. These entries will be used to pinpoint the timeline of learning activities. Students' final marks were added in a post-clustering analysis to explore correlations between the groups discovered and the academic performance.

Feature Extraction Results

To analyse the behavioural features described in the literature, this study focuses on two dimensions: learning tactics and activity level which serves as proximities of engagement levels (Matcha et al., 2020; Song et al., 2024). Learning tactics refer to the types of learning resources students interact with, including the number of discussion accesses (**F1**), file accesses (**F2**), page accesses (**F3**), and quiz accesses (**F4**). Activity level, on the other hand, captures when students interacted with the LMS, segmented by academic periods, during the teaching weeks (**F5**), mid-semester break (**F6**), SWOT-Vac and examination period (**F7**), and after the examination period (**F8**). These features represent what students do and when they do it, offering deep understanding of their learning patterns.

To capture students' preparedness and alignment with the flipped classroom design (Gonda et al., 2021), we use the approach in (Armas-Cervantes et al., 2023) where PM was used to extract features related to the number of interactions before the seminar (**F9**) and after the seminar (**F10**).

Students preferred learning hours (chronotype) was extracted based on the most frequent time of interaction (**F11**), as this was shown to correlate with academic performance (Enright & Refinetti, 2017). Finally, repetition (**F12**) was calculated by subtracting the number of unique interactions from the total interactions, based on evidence that repetition improves performance (Musfeld et al., 2023). The 12 features were extracted for each of the students in the dataset and the features were compiled in a feature table for further processing (see Table 1).

Table 1. Extracted features snapshot.

name_hash	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Student1	98	1	20	1	58	41	21	0	5	115	3	97
Student10	133	1	18	1	84	0	69	0	58	95	3	129
…	…	…	…	…	…	…	…	…	…	…	…	…
Student184	426	7	48	4	197	10	278	0	68	417	3	462
Student185	130	3	18	1	76	0	76	0	26	126	3	129

Clustering Results

Different clustering techniques and strategies were tested considering the high dimensionality of the data which can affect the clustering techniques. We compared two clustering techniques: DBSCAN and K-Means. Initial experiments considered the data with the 12 features for clustering. The results from DBSCAN and K-means are presented in Figure 2; these figures represent the clustering results in a 2-dimensional plane where each point represents a student. For visualisation purposes, the dimensionality reduction was done after the clustering using PCA. In the case of DBSCAN, a large proportion of data points were labelled as noise (blue dots in Figure 2, left). Similarly, K-Means was tested on all 12 features but consistently produced unbalanced clusters (clusters contained less than 10% of students) due to curse of dimensionality.

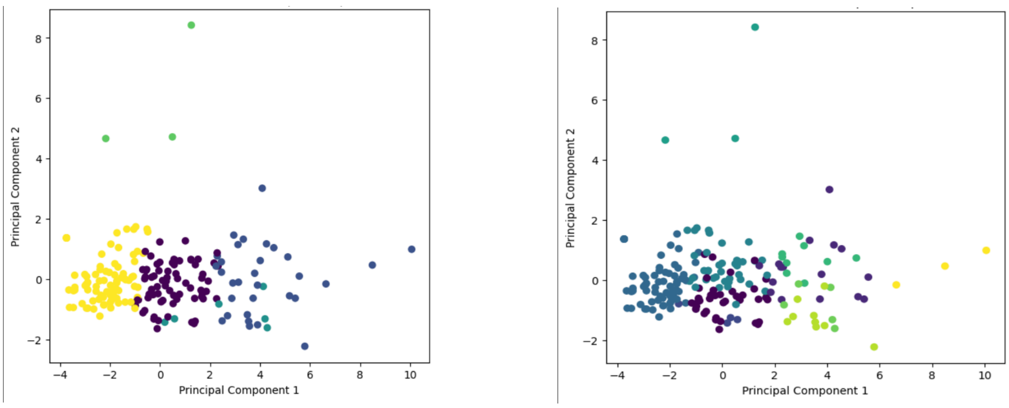

Figure 2. DBSCAN clustering result, with most data labelled as noise in blue (left), and K-Means iterations showing no clear separation between clusters (right).

To address high dimensionality and enhance interpretability, PCA was applied for dimensionality reduction before the clustering step. A scree plot (Figure 3, left) showed that only the first principal components captured substantial variance. Hence, using PCA, only the top two components were used before the clustering step.

Both DBSCAN and K-means were used to cluster the dimensionality-reduced data. However, DBSCAN labelled a large portion of data points as noise. Hence, K-means was selected as the clustering algorithm. We tested K-Means on the two principal components with K values ranging from 1 to 10 to see the different outcomes. The results showed a distinct separation between clusters. Since inappropriate K values can adversely affect precision clustering (Fahrudin, 2021), the Elbow Method was applied to determine the optimal number of clusters, which suggested a suitable K value based on the point of inflection in elbow curve (Figure 3, right).

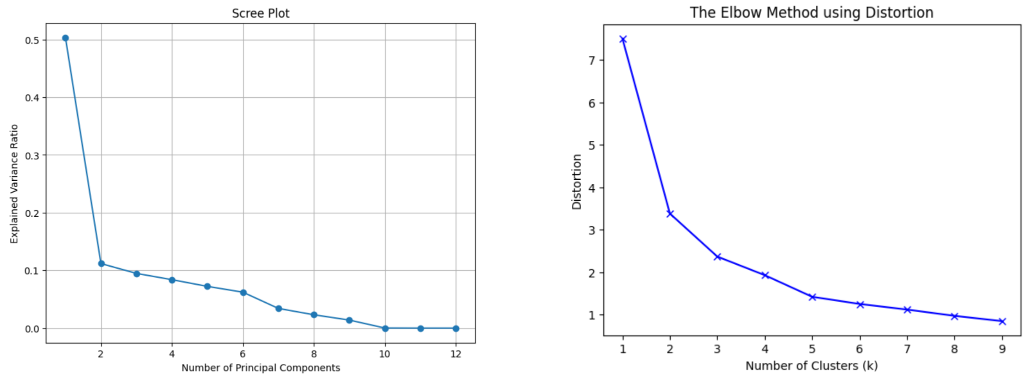

Figure 3. Scree plot shows Principal Component 1 explained about 50% variance (left). Elbow method showed elbow point at K=5 (right)

Using the Elbow Method, we chose a K = 5, where the distortion values begin to decrease more gradually (Figure 3, right). K-Means clustering on Principal Components 1 and 2 identified 5 distinct student groups with clear separation (Figure 4). Cluster labels were mapped back to the original feature table, and average values of the 12 behavioural features were calculated for each cluster (Table 2). Final subject marks were also integrated to explore correlations between behaviour and performance. While mark differences were not significant, clear patterns were observed. These insights formed the basis to generate personas representing behavioural traits and academic trends for each student profile.

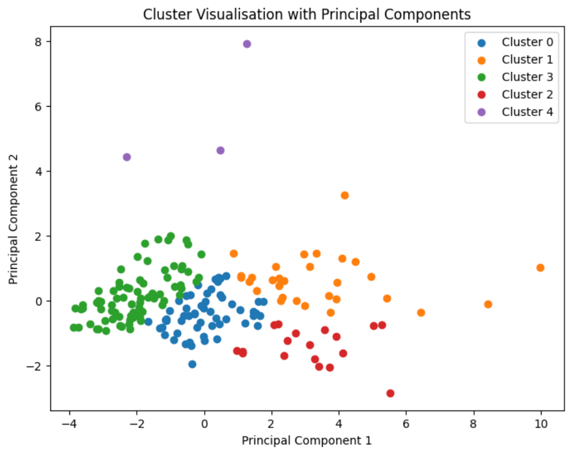

Figure 4. Visualisation for K-Means on 2 principal components (K=5). Cluster 4 behave differently from others (outliers).

Table 2. Statistical information snapshot (K=5, Decimal place rounded for better visibility).

Cluster	0	1	2	3	4
F1	110	208	205	47	91
F2	2	4	3	1	4
F3	16	26	25	8	16
F4	2	4	3	1	3
F5	80	118	165	31	37
…	…	…	…	....	…
F9	31	43	98	7	4
F10	98	198	137	51	109
F11	3	3	3	3	1
F12	108	218	212	43	94
Mark	67	70	71	65	61
Count	56	30	18	78	3

Persona Results

The five clusters represent groups of students with similar learning patterns according to the defined features. The final personas can be generated from statistical information in Table 2, where each cluster captured a distinct combination of learning behaviours shared by the respective students. Python Imaging Library (PIL) script was used to automatically generate high-level persona overviews by summarising the key statistics of each cluster. Predefined names were assigned to differentiate personas, and visual elements were added using icons from Flaticon.com (PIL does not support built-in icons).

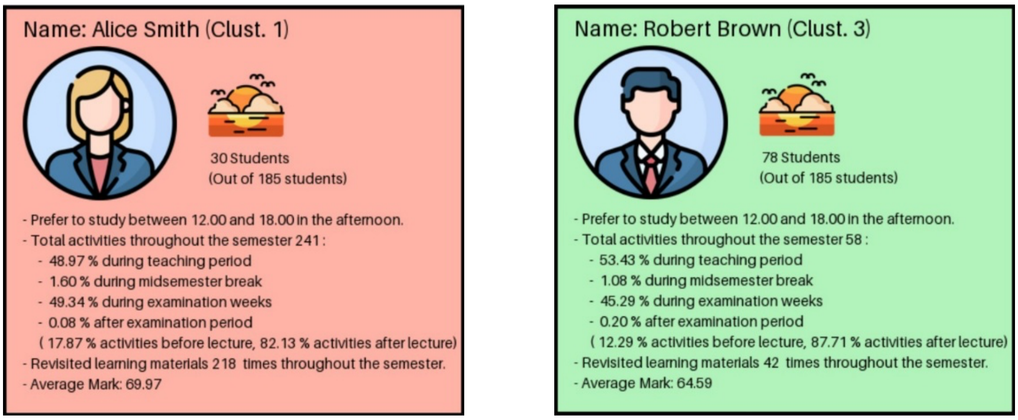

Figure 5. Automatically generated persona based on cluster-specific information (colours are for illustrative purpose and have no relation with persona characteristics).

An example of the automatically generated persona visualisation is shown in Figure 5. Due to space restrictions, only 2 of the 5 generated personas are shown. This figure illustrates clear differences in activity levels, learning strategies, pre-readings, repetition frequency, and chronotypes across the clusters. Although the differences in academic performance were not statistically significant, a consistent trend was observed across clusters.

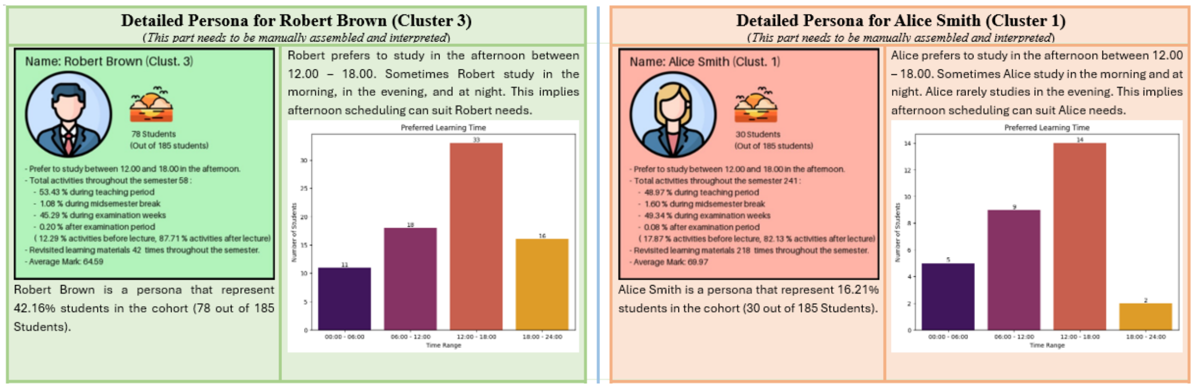

Figure 6. Example interpretation of chronotype for Cluster 3 and Cluster 1.

For deeper insight, these initial persona representations were manually interpreted and expanded into more descriptive narratives. In this study, Cluster 3 and 1 were chosen to represent passive and highly active students. The example of chronotype interpretations is presented in Figure 6. Please note that additional charts can be added to the personas according to the statistics of each feature in the cluster; however, due to space restrictions, we limited ourselves to show 1 chart.

Discussions

The results highlight the value of LMS event logs in revealing distinct learning profiles within a cohort, producing actionable insights for educators seeking to better understand student needs. In the analysed data, five personas were extracted from clustering the defined behavioural features, each reflecting various levels of activities, learning strategies, and temporal study habits. Notably, Cluster 1, representing highly active students (more than 202 interactions), consistently accessed learning materials before lecture dates and performed above the cohort average. This proactive behaviour aligns with the principles of flipped classrooms, where pre-lecture preparation enhances in-class learning and post-lecture consolidation. In contrast, Cluster 3, the least active group (fewer than 117 interactions), demonstrated passive interaction and achieved the lowest average final mark (64.59), suggesting that surface-level interaction is insufficient for strong academic outcomes.

Interestingly, Cluster 4 contained students with moderate interaction levels (around 113), yet underperformed (average mark 60.67), raising questions about factors beyond LMS interaction that may influence performance (e.g., personal issues, external workload, or assessment design). This indicates that while the defined features are informative, they do not capture the full picture of academic success.

Patterns in temporal activities also revealed key trends: engagement decreased during the mid-semester break due to the absence of new content but rose sharply during Weeks 10 - 11, likely as students prepared for final assessments. In Week 12 the activity levels drop likely due to assignment deadlines.

These findings show that LMS event logs can be meaningfully translated through personas to distinguish learning behaviour types in a scalable and automated way. By constructing personas, academics are provided with easy visual representation of their cohort, and behavioural study patterns of student clusters. These patterns provide some insight into early detection of students at-risk, thus be able to tailor support strategies and reflect on course design. For example, groups showing last-minute surges or low pre-lecture activity may benefit from targeted contents or special guidance. For instance, only 12.29% of the activities were done pre-lecture in Cluster 3, see Figure 5 (right), and there was low material repetition (42 times through the whole semester). Further, these patterns of behaviour could be shared with students as a way of motivating and directing their learning, reflecting on what material/resources they must access.

The presented results show the potential of integrating ML and PM for both macro-level cohort segmentation and micro-level activity tracing, demonstrating the complementary strengths of these approaches. However, some limitations remain, such as the exclusion of study duration, which was deemed unreliable due to data quality issues and limitation in the LMS logging system. Overall, the presented pipeline offers a replicable method for converting LMS logs into actionable insights through automated methods. These data-driven personas bridge the gap between raw interaction data and pedagogical decision-making, helping academics make informed choices about intervention timing, content delivery, and student support.

Conclusion & Future Directions

This research offers a proof of concept for integrating machine learning and process mining to understand student behaviours in education. The methodology proved effective by extracting features from LMS event logs and applying clustering to reveal interpretable behavioural patterns that characterise distinct student groups. Consistent with prior research (Matcha et al., 2020; Song et al., 2024), the correlation between learning styles and academic performance was not strongly significant, however patterns in learning tactics, activity levels, preparedness, and repetition showed relationships with academic performance, while chronotype exhibited no clear pattern.

The findings have practical implications for academics seeking to identify at-risk students and intervene early on. Subject coordinators can use persona analysis to tailor course design to satisfy student needs. Additionally, students may gain awareness of how their learning patterns impact academic success, potentially improving study motivation. Furthermore, our approach can be fully automated as data is automatically gathered by the LMS, hence it allows academics to generate personas at different points in time during the semester to capture changes in student behaviour.

Although the pipeline is reusable, the optimal number of personas may vary depending on the nature of the dataset. Interpretation of LMS interactions assumes all activities are individual and meaningful, potentially overlooking passive or collaborative behaviours. Finally, while this study relied solely on behavioural log data, incorporating qualitative inputs (e.g., interviews or surveys) in future research and other relevant features could provide deeper insights into learner motivations and challenges, enriching persona development and activity analysis.

References

Agha, D., Meghji, A. F., & Bhatti, S. (2023). Clusters of Success: Unpacking Academic Trends with K-Means Clustering in Education. *VFAST Transactions on Software Engineering*, *11*(4), 15–31. https://doi.org/10.21015/vtse.v11i4.1633

Armas-Cervantes, A., Ehsan, A., & Farbod, T. (2024). Dusting for fingerprints: Tracking online student engagement. *Computers and Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100232

Armas-Cervantes, A., Mendoza, A., & Abedin, E. (2023). Uncovering Students' Learning Pathways: A Process Mining Perspective. *34th Australasian Association for Engineering Education Conference*, 746–754. https://doi.org/10.3316/informit.T2024110500004191599089935

Brown, M. (2023, January 8). *Personas Are Living Documents: Design Them to Evolve*. https://www.nngroup.com/articles/personas-are-living-documents/

Enright, T., & Refinetti, R. (2017). Chronotype, class times, and academic achievement of university students. *Chronobiology International*, *34*(4), 445–450. https://doi.org/10.1080/07420528.2017.1281287

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Fahrudin, N. F. (2021). Sequence Clustering in Process Mining for Business Process Analysis Using K-Means. *Journal MIND Journal | ISSN*, *6*(1), 16–30. https://doi.org/10.26760/mindjournal.v6i1.16

Gonda, D., Pavlovičová, G., Tirpáková, A., & Ďuriš, V. (2021). Setting up a flipped classroom design to reduce student academic procrastination. *Sustainability (Switzerland)*, *13*(15). https://doi.org/10.3390/su13158668

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2019). *Rule-Based Evolutionary Online Learning Systems: A Principled Approach to LCS Analysis and Design*. Scholars Portal.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Liu, M. (2022). The Relationship between Students' Study Time and Academic Performance and its Practical Significance. In *BCP Education & Psychology ESS* (Vol. 2022).

Macqueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Symp. on Math. Statist. and Prob.*, 281–297.

Matcha, W., Gašević, D., Ahmad Uzir, N., Jovanović, J., Pardo, A., Lim, L., Maldonado-Mahauad, J., Gentili, S., Pérez-Sanagustín, M., & Tsai, Y. S. (2020). Analytics of learning strategies: Role of course design and delivery modality. *Journal of Learning Analytics*, *7*(2), 45–71. https://doi.org/10.18608/JLA.2020.72.3

Musfeld, P., Souza, A. S., & Oberauer, K. (2023). Repetition learning is neither a continuous nor an implicit process. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(16). https://doi.org/10.1073/pnas.2218042120

Park, D.-H., Yu, J.-H., & Yoo, H.-Y. (2024). The Case Study on Development of Segmentation and Data-driven Persona Based on OTT Service Usage Logs: Focusing on Netflix. *Journal of the Korea Academia-Industrial Cooperation Society*, *25*(3), 312–325. https://doi.org/10.5762/kais.2024.25.3.312

Song, Y., Oliveira, E., Kirley, M., & Thompson, P. (2024). A Case Study on University Student Online Learning Patterns Across Multidisciplinary Subjects. *ACM International Conference Proceeding Series*, 936–942. https://doi.org/10.1145/3636555.3636939

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). https://doi.org/10.1186/s41239-021-00313-7

van der Aalst, W. (2012). Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems*, *3*(2). https://doi.org/10.1145/2229156.2229157

Copyright statement