Santos, Ricardo; Henriques, Roberto

Decoding student success in higher education : a comparative study on learning strategies of undergraduate and graduate students

Studia paedagogica. 2023, vol. 28, iss. 3, pp. [59]-87

ISSN 1803-7437 (print); ISSN 2336-4521 (online)

Stable URL (DOI): https://doi.org/10.5817/SP2023-3-3
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.79679
Access Date: 29. 03. 2024
Version: 20240320

STUDIA PAEDAGOGICA

vol. 28, n. 3, 2023
https://doi.org/10.5817/SP2023-3-3

DECODING STUDENT SUCCESS IN HIGHER EDUCATION: A COMPARATIVE STUDY ON LEARNING STRATEGIES OF UNDERGRADUATE AND GRADUATE STUDENTS

Ricardo Santos[a], Roberto Henriques[a]

[a] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal

ABSTRACT

Learning management systems (LMS) provide a rich source of data about the engagement of students with courses and their materials that tends to be underutilized in practice. In this paper, we use data collected from the LMS to uncover learning strategies adopted by students and compare their effectiveness. Starting from a sample of over 11,000 enrollments at a Portuguese information management school, we extracted features indicative of self-regulated learning (SRL) behavior from the associated interactions. Then, we employed an unsupervised machine learning algorithm (k-means) to group students according to the similarity of their patterns of interaction. This process was conducted separately for undergraduate and graduate students. Our analysis uncovered five distinct learning strategy profiles at both the undergraduate and graduate levels: 1) active, prolonged and frequent engagement; 2) mildly frequent and task-focused engagement; 3) mildly frequent, mild activity in short sessions engagement; 4) likely procrastinators; and 5) inactive. Mapping strategies with the students' final grades, we found that students at both levels who accessed the LMS early and frequently had better outcomes. Conversely, students who exhibited procrastinating behavior had worse end-of-course grades. Interestingly, the relative effectiveness of the various learning strategies was consistent across instruction levels. Despite the LMS offering an incomplete and partial view of the learning processes students employ, these findings suggest potentially generalizable relationships between online student behaviors and learning outcomes. While further validation with new data is necessary, these connections between online behaviors and performance could guide the development of personalized, adaptive learning experiences.

KEYWORDS

self-regulated learning; student strategies; learning management systems; higher education; machine learning

CORRESPONDING AUTHOR
Ricardo Santos, NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, 1070-312, Lisbon, Portugal
e-mail: rcsantos@novaims.unl.pt

Introduction

In the evolving landscape of education, the focus has shifted toward individual student progress, significantly altering the dynamics of teaching and learning. This transformation is largely driven by the advent of artificial intelligence tools, leading to substantial investments in personalized learning and intelligent tutoring systems (Holmes & Tuomi, 2022). Despite these advancements, traditional educational methods continue to hold relevance, even as educators grapple with challenges such as larger class sizes and the rise of remote learning, thus reducing the reliability of conventional ways of assessing student progress, such as attendance and in-class behavior (Bellur et al., 2015). These changes challenge educators to identify and support students who require the most assistance.

Consequently, there is a growing emphasis on self-regulated learning (SRL) behaviors, which provide a more comprehensive insight into a student's abilities, motivations, and attitudes toward learning. SRL skills are crucial for students, particularly in higher education where autonomy is expected (Boekaerts, 1997; Broadbent & Poon, 2015) and in the 21st century workplace, where employers prioritize learners who can take charge of their development (Trilling & Fadel, 2009). Thus, gauging and fostering the development of SRL behavior is imperative in both educational and professional settings.

The cyclical model of SRL involves students actively participating in their learning through cycles of forethought, performance control, and self-reflection (Zimmerman, 2000, 2002). Students develop tools to regulate their cognition, behavior, and emotions through repeated engagement in these processes (Zimmerman & Moylan, 2009). A key component of SRL is the development of strategies that enhance students' ability to achieve their learning goals. According to Pintrich et al. (1991), SRL strategies can be divided into three categories: cognitive, metacognitive, and resource management. Both time management and effort regulation are positively correlated with student performance (Broadbent, 2017; Puzziferro, 2008). Conversely, evidence suggests that students with underdeveloped SRL behaviors struggle in contexts where more autonomy is expected, such as in online and blended learning contexts (Broadbent, 2017).

One approach to measure SRL behaviors is direct observation of the students. For example, timing how long it takes for a student to finish a set

of tasks provides behavioral evidence of the SRL trait of time management (Winne & Jamieson-Noel, 2002). However, comprehensively observing student learning behaviors through direct means can be difficult in practice, as designing rigorous experiments in controlled settings requires extensive time and resources (Susac et al., 2014). Alternatively, self-report questionnaires, such as the Motivated Strategies for Learning Questionnaire (MSLQ), allow students to self-evaluate SRL traits (Pintrich et al., 1991; Winne & Perry, 2000). These questionnaires are inexpensive and simple to administer, but sole dependence on student self-reports poses risks of bias and only reflects students' perceptions at the time of administration.

In recent years, the widespread adoption of learning management systems (LMS) in higher education institutions has increased the availability of detailed student trace data (Coates et al., 2005). These systems record students' digital interactions within their learning environment. By applying data mining techniques to these logs, researchers can extract variables (from this point onward referred to as features) connected to SRL behaviors (Baker et al., 2020). These features can be used to gain additional insights about learners, the learning processes they engage in, and their academic progress. For example, supervised machine learning algorithms have been successful at flagging students at risk of failing (Bernacki et al., 2020; Macfadyen & Dawson, 2010; Riestra-González et al., 2021). Alternatively, unsupervised machine learning algorithms (also referred to as clustering algorithms) can be used to uncover learner strategy profiles (Cerezo et al., 2016; Riestra-González et al., 2021).

While prior works have utilized unsupervised machine learning to identify learning behaviors from LMS data, a limited number of studies apply these approaches, especially for large, multi-course samples. Moreover, exploring possible differences and effectiveness of learning strategies across different instruction levels is still a relatively unexplored topic. This work aims to address these gaps by leveraging clickstream data to extract course-agnostic features from an LMS, identify learner strategy profiles at the undergraduate and graduate levels, and assess their relative effectiveness for academic success. The research questions are:

1. What course-agnostic learning strategy profiles can be extracted from undergraduate and graduate students' SRL features extracted from LMS data?
2. What is the relationship between the learning strategies uncovered by *k*-means and end-of-course performance at each instruction level?
3. Are there differences in the effectiveness of the learning strategies between instruction levels?

To answer these questions, Moodle logs were collected from 57 undergraduate and 124 graduate courses taught at a Portuguese information management

school during the 2020/2021 academic year. From these logs, 30 SRL features were extracted to build a dataset, which was then split between undergraduate and graduate course enrollments. The *k*-means clustering algorithm was used to identify learner strategy profiles at each instruction level, allowing the comparison of the effectiveness of each strategy.

The remainder of this paper is structured as follows: The next section provides an overview of prior research utilizing unsupervised learning approaches to identify learner strategy profiles from LMS data. The third section presents the study's data and methodology. The fourth section presents the results. The fifth section discusses the results, their alignment with expected outcomes, and key implications. The sixth and final section concludes with a summary of the main findings and a discussion of future research directions.

1 Related work

This section provides an overview of research that uses unsupervised machine learning techniques to identify learning strategies from SRL-related features. The main purpose of this section is to discuss the different existing approaches regarding the adoption of theoretical frameworks, sample size, features extracted, the techniques used and the author's main finding when uncovering learning strategies from data. A literature review table featuring all works covered in this section is provided in Table 1.

The theoretical frameworks most frequently cited include Biggs' 3P model (Biggs, 1987) and the SRL motivational model of Pintrich et al. (1991). These theoretical foundations provide a clear interpretive lens for variables derived from LMS data, a solid rationale for the chosen tools, and a frame of reference for interpreting results. For example, Gašević et al. (2017) used the Study Process Questionnaire (SPQ) instrument to supplement LMS data, which enabled them to distinguish between deep and surface learning indicators among their students. They discovered that students who employ deep learning strategies outperform their peers. Li & Tsai (2017) also reported using the MSLQ to uncover SRL variables from their students to map SRL to academic performance. However, most studies reviewed do not delve extensively into a theoretical SRL framework (Cerezo et al., 2016; Moubayed et al., 2020; Riestra-González et al., 2021). Instead, they merely reference existing frameworks to rationalize how LMS data can reveal learning strategies and the reasons behind selecting specific feature types. This trend could be attributed to a greater focus on using these variables to uncover learning strategy profiles from data (Cerezo et al., 2016; Riestra-González et al., 2021) rather than conducting a thorough discussion of how a specific SRL model explains differences in academic performance or achievement. Another potential reason stems from the nature of the data used.

Table 1
Literature review table research on the use of unsupervised learning to uncover learning strategies

Reference	Grounding	Sample	Source	Variables	Clusters	Key Findings
Hung & Zhang, (2008)	Not specified	98 undergrad students	LMS	5 engagement features (e.g. frequency, materials)	*k*-means (3 clusters)	Active students performed better academically
Cerezo et al. (2016)	Not specified	140 undergrad psychology students	LMS	Interactions with LMS materials	*k*-means (5 clusters)	Better performers procrastinate less while also spending less time on the LMS
Gašević et al. (2017)	Biggs' 3P model	144 undergrad students	LMS, survey	Deep/surface learning indicators from LMS	Hierarchical clustering (2 clusters)	Deep learning strategies map to deep learning scales on survey
Li & Tsai (2017)	Pintrich's motivational model	59 undergrad computer science students	LMS, survey	Time on materials, SRL features	*k*-means (3 clusters)	Consistent use students had higher motivation and achievement
Çebi & Güyer (2020)	Pintrich's motivational model	122 undergrad statistics students	LMS	Time on activities	Unspecified clustering (3 clusters)	Engagement associated with better performance and motivation
Matcha et al. (2020)	Biggs' 3P model	~1400 students in a MOOC	LMS	Video, quiz, assignment, forum actions	Hierarchical clustering & Markov model (4 clusters)	Active/highly active had better performance
Moubayed et al. (2020)	None specified	486 undergrad students	LMS	Time on activities, assignments	*k*-means (3 clusters)	Highly engaged students tend to perform better
Yang et al. (2020)	None specified	242 undergrad students	LMS	Homework submission patterns	*k*-means (3 clusters)	Non-procrastinators perform better
Riestra-González et al. (2021)	None specified	~16000 students in 699 courses	LMS	31 features (resource accesses)	*k*-means (6 clusters)	More engaged students perform better

While LMS data is rich, it is essentially a series of timestamped actions. Features like click count can be categorized under Pintrich et al.'s (1991) resource management, but they only offer a partial and indirect insight into crucial constructs such as motivation or emotional state, which are prevalent in popular SRL models (Panadero, 2017).

The sample sizes used in these works also differ greatly. They range from a small group of 59 students in a single course (Li & Tsai, 2017) to a large cohort of nearly 16,000 students spread across 699 different courses (Riestra-González et al., 2021). This significant variation limits the ability to derive insights that can be generalized across different contexts. Moreover, although the LMS is a common data source in the reviewed works, the specific features and contexts for variable usage and extraction vary substantially. Several studies track the frequency of specific student actions or the time spent on the LMS (Cerezo et al., 2016; Matcha et al., 2020; Riestra-González et al., 2021). However, different sets of features have been extracted from the LMS, with Yang et al.'s (2020) approach using the LMS to extract and analyze features related to procrastination behaviors on homework deadlines.

In the process of uncovering learning strategies, the most common method is to group students into clusters using k-means or hierarchical clustering algorithms based on the features extracted from the LMS logs (Cerezo et al., 2016; Hung & Zhang, 2008; Moubayed et al., 2020). For example, Hung & Zhang (2008) extracted five LMS engagement features from 98 students in an online course and used k-means to uncover three clusters that differentiated poor-performing versus above-average students. Similarly, Cerezo et al. (2016) identified four learner strategy profiles in a sample of 140 students using k-means on LMS trace data, finding the cluster with socially-focused and strategic study habits achieved the highest grades. Riestra-González et al. (2021) also found significant differences in four out of the six learning strategies uncovered. Beyond k-means, both Gašević et al. (2017) and Matcha et al. (2020) used hierarchical clustering to group similar sets of students. Finally, Çebi & Güyer (2020) did not mention the specific algorithm used in their work despite also using a clustering technique to uncover three distinct learning strategy profiles from LMS data.

Observations from multiple studies have consistently shown that students who exhibit higher engagement and less procrastination tend to achieve better academic results than their peers (Cerezo et al., 2016; Moubayed et al., 2020; Yang et al., 2020). Tactics such as evenly spacing study time and completing assessments early were positively associated with achievement, while students exhibiting low numbers of clicks, and late and infrequent logins tended to perform worse (Hung & Zhang, 2008; Li & Tsai, 2017; Matcha et al., 2020). These findings align with expectations, as students who demonstrate traits related to the employment of an actual strategy are more likely to have more

developed SRL skills. However, only a few studies mapped clusters directly back to established SRL frameworks to confirm theoretical connections between engagement and motivation (Çebi & Güyer, 2020; Li & Tsai, 2017). In terms of implications, the findings of these studies point to the potential of analytics tools that aim to provide adaptive interventions and personalized support starting from the student behaviors (Cerezo et al., 2016).

The research discussed in this section illustrates that using unsupervised machine learning techniques to uncover students' learning strategies with LMS data is an active and growing area of study. A common approach is to use clustering algorithms to group students based on their interactions with course materials and activities. These clusters are then associated with academic performance metrics or self-reported surveys to draw connections between learning strategies, motivation, and achievement.

However, there are notable gaps worth highlighting. Small sample sizes are a common issue, and no studies have explicitly sought to identify and compare learning strategies across different levels of instruction. This limits the generalizability of findings and hinders the development of comprehensive models. Additionally, there are inconsistencies in the features used by different authors, partly due to the absence of a consistent theoretical framework for SRL in most works. This leads to disparate findings and interpretations. While addressing this gap is beyond the scope of this work, adopting a robust theoretical framework could lead to more consistent and comparable findings across studies.

This work aims to address some research opportunities by using larger samples and more courses, contributing to more generalizable models. This could help determine if students' learning strategies can be replicated in a general context and inform the design of personalized learning experiences on the LMS, potentially reducing student dropout rates and improving achievement.

2 Methodology

This work started with the extraction of anonymized institutional Moodle logs and their transformation into a structured dataset indexed by program, course, and student, accompanied by 30 features associated with the resource management construct found in Pintrich's motivational model for SRL (Pintrich et al., 1991). The dataset was split into undergraduate and graduate subsets and given to separate instances of the *k-means* clustering algorithm (Macqueen, 1967). The resulting clusters were characterized and compared. A summary of the adopted approach is depicted in Figure 1. Unless otherwise noted, all data manipulation and analysis procedures were implemented using Python (McKinney, 2017) and Scikit-learn (Pedregosa et al., 2011).

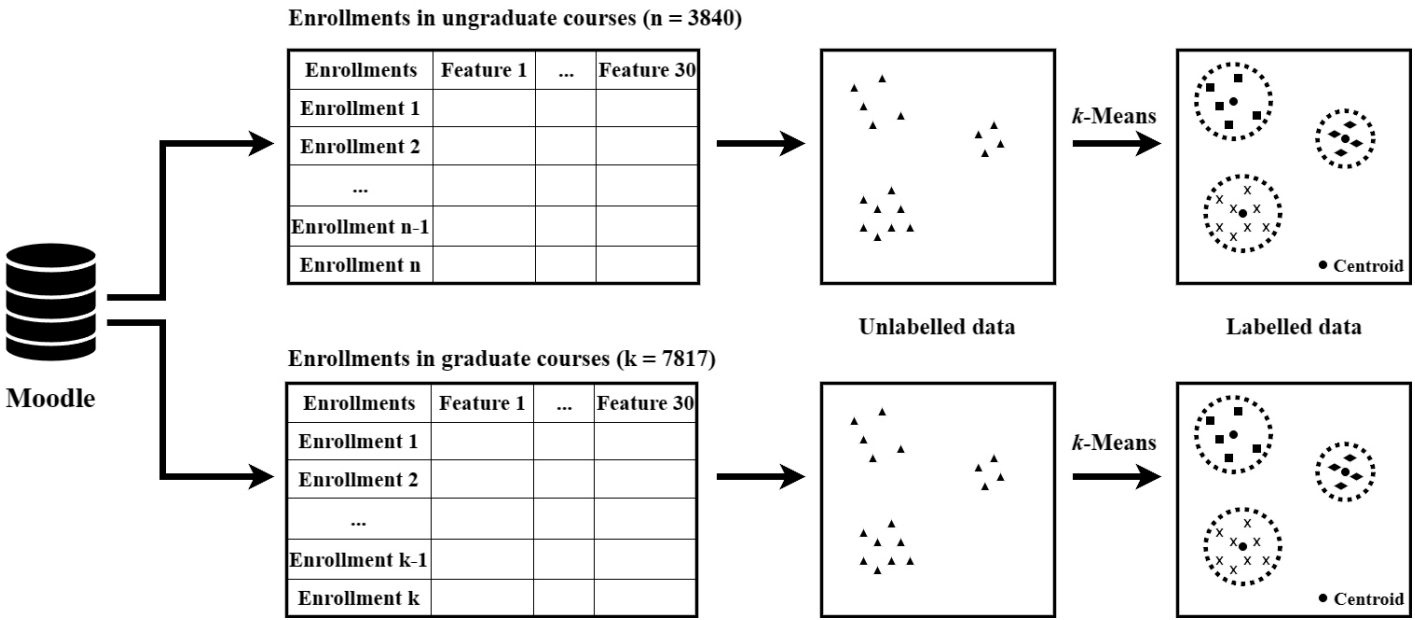

Figure 1
Overview of the approach

2.1 Data

The data was collected from a Portuguese information management school in 2020/2021, which offers graduate and undergraduate programs in data science, information management, and information systems and technologies. The sample includes 1564 graduate and 409 undergraduate students enrolled in 124 and 57 courses, respectively, totaling 11,297 student enrollments. Moodle logs and end-of-course final grades were accessed for each enrollment with no additional data sources being considered. Table 2 presents an overview of the population for each instruction level, including the number of courses, students, enrollments, and average end-of-course performance, which in the Portuguese systems assumes values between 0 and 20, with 10 representing the minimum passing threshold. All student data was anonymized in compliance with GDPR, and the project was approved by the *Ethics Committee and Institutional Review Board* with Code DSCI2022-9-227363.

Table 2
Summary of the characteristics of courses and students per instruction level (grades ranging from 0 to 20)

	Courses	Students	Enrollments	Enrollments per course	Average end-of-course grade (± Standard Deviation)
Undergraduate					
Program A	28	160	1336	47.71	13.48 ± 4.30
Program B	29	249	2144	73.93	14.22 ± 4.02
Sub-total	**57**	**409**	**3480**	**63.87**	**13.94 ± 4.14**
Graduate					
Program 1	10	33	322	32.20	13.68 ± 3.23
Program 2	17	173	755	44.41	13.98 ± 3.52
Program 3	6	31	170	28.33	15.78 ± 2.20
Program 4	6	40	218	36.33	15.10 ± 2.51
Program 5	4	310	120	30.00	16.27 ± 2.36
Program 6	4	27	108	27.00	16.76 ± 1.15
Program 7	9	155	666	74.00	16.50 ± 2.72
Program 8	13	36	391	30.08	15.74 ± 2.68
Program 9	7	82	267	38.14	16.48 ± 1.65
Program 10	2	33	54	27.00	13.07 ± 5.39
Program 11	15	192	1818	121.20	15.73 ± 3.07
Program 12	29	416	2857	98.52	15.67 ± 3.06
Program 13	2	36	71	35.50	15.87 ± 2.65
Sub-total	**124**	**1564**	**7817**	**81.89**	**15.54 ± 3.08**
Total	**181**	**1973**	**11,297**	**72.88**	**15.04 ± 3.52**

2.2 Feature Extraction

The first part of the process involved converting Moodle logs into data structures suitable for statistical analysis. For each course, Moodle keeps a timestamped record of every click made on the LMS, including which student made the click and where it was performed within the LMS. To extract meaningful features from this data, we adopted three perspectives that measure student engagement with the LMS, a critical resource for our students: *Raw activity*, which refers to the number of times a certain action is performed; *Time-on-task*, which refers to the amount of time dedicated to studying on LMS; and *Procrastination*, which measures at which stages of the course the students log into the LMS.

In total, 30 candidate features were extracted and considered for subsequent steps. The reasons for the choice of these specific features are two-fold. First, these features fall under the resource management construct of Pintrich's motivational model for SRL (Panadero, 2017) and measure student interaction with the LMS. Moreover, these features have also been successfully utilized in a plethora of previous learning analytics research (Aljohani et al., 2019; Conijn et al., 2017; Riestra-González et al., 2021; Romero et al., 2013; Santos & Henriques, 2023). Table 3 provides a comprehensive list of the features extracted from the logs and their respective averages and standard deviations for each instruction level.

Table 3
Extracted candidate features

Feature	**N (undergraduate)**	**Mean ± Standard Deviation (undergraduate)**	**N (graduate)**	**Mean ± Standard Deviation (graduate)**
Perspective 1: Raw activity				
Total clicks (n)	3480	279.59 ± 177.69	7817	248.56 ± 168.54
Clicks (% of course total)	3480	1.64 ± 1.01	7817	1.59 ± 1.60
Forum clicks (n)	2222	8.63 ± 13.85	6390	19.62 ± 25.31
Forum posts (n)	27	1.52 ± 1.08	375	1.76 ± 1.13
Discussions viewed (n)	1283	6.47 ± 8.98	5277	11.38 ± 14.47
Folder clicks (n)	1825	20.20 ± 28.40	63.68	20.29 ± 23.64
Resources viewed (n)	3460	64.16 ± 55.39	6928	46.96 ± 38.85
URLs viewed (n)	2419	25.14 ± 16.96	5514	17.65 ± 13.67
Course clicks (n)	3480	110.48 ± 80.62	7816	94.41 ± 66.97
Assessments started (n)	1688	3.02 ± 2.56	3763	2.56 ± 3.16
Assignments viewed (n)	974	17.40 ± 24.88	3618	12.98 ± 17.51

Assignments submitted (n)	893	5.55 ± 5.81	3113	4.90 ± 4.83
Submissions (% of course total)	893	3.36 ± 2.56	3113	2.60 ± 5.14
Perspective 2: Time-on-task				
Online sessions (n)	3480	59.96 ± 41.77	7817	47.92 ± 29.82
Clicks/session (n)	3480	4.96 ± 2.33	7817	5.40 ± 2.80
Clicks/day (n)	3480	1.85 ± 1.18	7817	2.02 ± 1.43
Total time online (min)	3480	491.71 ± 394.82	7817	396.98 ± 331.03
Aver. duration of online sessions (min)	3480	8.06 ± 3.85	7817	8.20 ± 5.64
Perspective 3: Procrastination				
Largest period of inactivity (h)	3480	463.88 ± 283.71	7817	415.60 ± 269.13
Days with 0 clicks (% of period)	3480	62.95 ± 11.87	7817	63.92 ± 11.88
PercCourse_1Login	3480	7.06 ± 9.32	7817	0.61 ± 8.52
PercCourse_NLogin ($n \in [2, 9]$)	…	…	…	…
PercCourse_10Login	3387	22.16 ± 15.10	7538	22.44 ± 19.45

2.3 Data analysis

The data for graduate and undergraduate students were processed separately but followed similar pipelines for preprocessing, feature selection, and clustering. The preprocessing stage involved three main steps. In the first step, the Jarque-Bera normality test (Jarque & Bera, 1980) was used to assess how reasonable it would be to assume the normal distribution of the data. This test measures the skewness and kurtosis of a feature and determines if it deviates significantly from those of a normal distribution (skewness of 0 and kurtosis of 3). In the second step, all features that could not be reasonably assumed to follow a normal distribution were transformed using the Yeo-Johnson power transformation (Yeo & Johnson, 2000). This method aims to transform non-normally distributed data into a shape resembling a normal distribution by raising the data to an appropriate power. The transformed variables were then standardized, which is the final step of the preprocessing stage. This rigorous preprocessing ensures that the data is appropriately conditioned for the subsequent stages of feature selection and clustering.

The feature selection process aimed to eliminate any variable that could be considered irrelevant or redundant for cluster construction from each perspective. This was achieved through a two-step strategy. The first step involved setting an absolute value of 0.8 on the Spearman correlation index

to flag potentially redundant variables. In the second step, *k*-means was used to create clustering solutions for each perspective, and the explained variance of each feature toward that solution was measured. Variables with very low explained variance (i.e. irrelevant variables) were removed, as were redundant features that exhibited the lowest explained variance. This process was repeated until a satisfactory clustering solution was achieved for each perspective. The resulting variables were then combined into a final dataset. Consequently, at the end of this stage, there were two preprocessed datasets: one containing the features necessary to build clusters on undergraduate enrollments, and another containing the features deemed relevant for clustering graduate enrollments.

In the third stage, each dataset was used as input to a separate instance of the *k*-means clustering algorithm. *k*-means is an iterative algorithm that groups data points based on distance, minimizing within-group distance while maximizing between-group distances. A key component of *k*-means is the concept of a centroid, which can be understood as a data point representing the coordinates of the center a group. By comparing the positions of these centroids, it is possible to understand the differences and similarities between the groups. Despite its simplicity, *k*-means enjoys widespread adoption when partitioning data into different groups (Wu et al., 2008). However, a limitation of *k*-means is that the number of resulting groups must be set *a priori*. In this implementation, the optimal number of groups (each referring to a learning strategy) was determined using the elbow method (Cerezo et al., 2016; Riestra-González et al., 2021) and found to be five for both instruction levels.

Once the groups were formed, they were analyzed to answer the research questions. To answer the first research question, the different learning strategies were characterized. This involved comparing the strategies adopted by students at the same instruction level to ensure there were no overlaps. The differences between learning strategies were measured by comparing the coordinates of the centroids determined by *k*-means. Between-group comparisons were performed at the feature level but interpreted at the perspective level. Two learning strategies were considered significantly different in one perspective if there were statistically significant differences in most variables belonging to that perspective. Due to the differences in scale, these comparisons were performed using standardized scores (0 mean and unit variance).

To answer the second research question, the average end-of-course grade associated with each learning strategy was calculated. This was followed by a comparison of the end-of-course grade of the various learning strategies at the same instruction level using Welch's t-test.

To answer the third and final research question, we performed a qualitative comparison of the learning strategies adopted by undergraduate students with those adopted by graduate students. The aim was to identify whether there were unique undergraduate or graduate-level strategies that did not exist at the other level of instruction. Moreover, the comparison also aimed to identify whether the relative effectiveness of strategies varied between the two instruction levels.

3 Results

3.1 Learning strategies in undergraduate and graduate students

The centroid coordinates presented in Table 4 show that all five resulting learning strategies differ significantly from one another regarding the *Raw activity* and *Time-on-task* perspectives, with strategies B and E not being significantly different when it comes to *Procrastination*.

From the perspective of *Raw activity*, students adopting different strategies exhibited varying levels of engagement with Moodle. Strategy D students exhibited the highest overall levels of engagement, with the highest number of clicks, both overall and across multiple pages, including resources, external links, and course page visits. Strategy C students had the second highest average engagement across most raw activity features, ranking highest in folder clicks and assessments started. In contrast, Strategy E students displayed the lowest raw activity engagement, with the least clicks across all features measured. Strategy A engagement was also relatively low, with all raw activity metrics falling below or slightly above average. Finally, while generally a low activity strategy, Strategy B students completed a relatively high number of assessment starts compared to other low engagement strategies.

Table 4

K-means standardized mean ± standard deviation for all variables in clustering for undergraduate enrollments (values with statistically significant (p-value < 0.05 on t-test) differences against all other strategies are in bold)

Feature	Strategy A (n =883)	Strategy B (n=615)	Strategy C (n=735)	Strategy D (n=702)	Strategy E (n=545)
Perspective 1: Raw activity					
Total clicks (n)	**−0.28 ± 0.39**	**−0.41 ± 0.49**	**0.79 ± 0.62**	**1.34 ± 0.84**	**−1.60 ± 0.53**
Folder clicks (n)	−0.29 ± 0.92	**0.05 ± 0.92**	**0.93 ± 0.77**	−0.25 ± 0.94	**−0.51 ± 0.71**
Resources viewed (n)	**0.08 ± 0.76**	**−0.47 ± 0.78**	**−0.07 ± 0.86**	**1.23 ± 0.84**	**−0.98 ± 0.79**
URLs viewed (n)	−0.04 ± 0.97	−0.21 ± 0.86	−0.12 ± 1.05	**0.85 ± 0.74**	**−0.68 ± 0.68**
Course clicks (n)	**0.02 ± 0.51**	**−0.67 ± 0.57**	**0.38 ± 0.57**	**1.45 ± 0.92**	**−1.49 ± 0.68**
Assessments started (n)	**−0.70 ± 0.53**	0.29 ± 0.84	**0.96 ± 0.65**	0.38 ± 1.10	**−0.88 ± 0.24**
Perspective 2: Time-on-task					
Online sessions (n)	**0.15 ± 0.48**	**−0.87 ± 0.55**	**0.31 ± 0.47**	**1.47 ± 0.97**	**−1.42 ± 0.72**
Clicks/session (n)	**−0.60 ± 0.66**	**0.85 ± 0.82**	**0.74 ± 0.74**	**−0.05 ± 0.88**	**−0.86 ± 1.02**
Clicks/day (n)	**−0.29 ± 0.41**	**−0.42 ± 0.50**	**0.80 ± 0.60**	**1.32 ± 0.78**	**−1.58 ± 0.46**
Total time online (min)	**−0.19 ± 0.53**	**−0.44 ± 0.48**	**0.56 ± 0.51**	**1.44 ± 0.87**	**−1.57 ± 0.53**
Average duration of online sessions (min)	**−0.39 ± 0.72**	0.58 ± 1.06	0.54 ± 0.74	**0.43 ± 0.83**	**−1.19 ± 0.90**
Perspective 3: Procrastination					
Largest period of inactivity (h)	**−0.22 ± 0.76**	0.85 ± 1.08	**−0.14 ± 0.81**	**−0.95 ± 0.96**	0.67 ± 2.16
Days with 0 clicks (% of period)	**−0.28 ± 0.64**	1.07 ± 0.62	**−0.13 ± 0.62**	**−1.24 ± 0.76**	1.01 ± 0.82
PercCourse_1Login	0.09 ± 1.05	0.09 ± 0.97	−0.19 ± 0.81	−0.20 ± 1.03	**0.26 ± 1.13**

Similar trends were observable for the *Time-on-task* and *Procrastination* perspectives, with some key exceptions. Aligned with their raw activity totals, Strategy D students spent the most time on the LMS, logged the highest number of sessions, and started accessing the system as early as possible, displaying low procrastination tendencies. Mirroring their overall inactivity, Strategy E students spent the least amount of time on the LMS, had the fewest sessions, and tended to start accessing the system later than the others. Strategies A and C again fell in between. Finally, the behavior displayed by students who adopted Strategy B was somewhat different. Their values on features related to *Procrastination* showed that they displayed values that were statistically similar to the highly inactive Strategy E students. However, there were some divergences between the *Raw activity* and *Time-on-task* perspectives as these students exhibited long sessions and the highest number of clicks per session of all learning strategies uncovered for undergraduate enrollments.

To facilitate interpretation, the strategies were labeled based on these engagement characteristics. Strategy A was termed *mildly frequent, mild activity in short sessions*, Strategy B *likely procrastinators*, Strategy C *mildly frequent and task-focused*, Strategy D *active, prolonged and frequent* and Strategy E *inactive*.

Table 5 presents the centroid coordinates for the five learning strategies uncovered for graduate students. A key difference between undergraduate and graduate enrollments is that forum clicks and assessments viewed impacted cluster construction for graduate students when they had provided little explanatory power for undergraduates. All five graduate learning strategies show significant differences across all perspectives.

From the perspective of *Raw activity,* students adopting Strategy 5 were the most engaged with Moodle materials, presenting the highest values for total clicks, clicks on course-related and resource pages, and assessments viewed. In contrast, students adopting Strategy 1 had the lowest levels of engagement across most features. The remaining strategies presented engagement values somewhere in between: Strategy 2 tended toward higher levels of engagement on most features; Strategy 4 tended toward lower values for total clicks but had high values for clicks on resources, external URLs, and assessment views; and Strategy 3 had close to average total clicks with high values for folder clicks and assessments started.

Table 5

K-means standardized mean ± standard deviation for all variables used in clustering of graduate enrollments (values with statistically significant (p-value < 0.05 on t-test) differences against all other groups are in bold)

Feature	**Strategy 1 (n =1381)**	**Strategy 2 (n=1697)**	**Strategy 3 (n=1503)**	**Strategy 4 (n=2093)**	**Strategy 5 (n=1143)**
Perspective 1: Raw activity					
Total clicks (n)	**−1.55 ± 0.55**	**1.02 ± 0.73**	**−0.02 ± 0.58**	**−0.31 ± 0.41**	**1.18 ± 0.93**
Folder clicks (n)	**−0.63 ± 0.80**	**0.66 ± 0.84**	**0.18 ± 0.88**	**−0.28 ± 0.97**	**0.04 ± 1.04**
Resources viewed (n)	**−0.94 ± 0.69**	0.16 ± 1.07	**−0.36 ± 0.80**	0.13 ± 0.75	**1.06 ± 0.85**
URLs viewed (n)	**−0.80 ± 0.36**	0.23 ± 0.73	**−0.35 ± 0.69**	0.23 ± 0.61	0.49 ± 0.99
Course clicks (n)	**−1.48 ± 0.69**	**0.79 ± 0.60**	**−0.41 ± 0.53**	**−0.11 ± 0.52**	**1.37 ± 0.92**
Assessments started (n)	**−0.84 ± 0.36**	0.87 ± 0.74	0.87 ± 0.69	**−0.67 ± 0.61**	**−0.08 ± 0.99**
Assessments viewed (n)	−0.51 ± 0.72	**−0.26 ± 0.85**	−0.52 ± 0.67	**0.43 ± 1.02**	**0.83 ± 0.93**
Forum clicks	**−0.41 ± 1.01**	0.54 ± 0.89	**−0.15 ± 0.97**	**−0.32 ± 0.87**	0.49 ± 0.91
Perspective 2: Time-on-task					
Online sessions (n)	**−1.35 ± 0.79**	**0.65 ± 0.56**	**−0.59 ± 0.55**	**−0.01 ± 0.55**	**1.40 ± 0.87**
Clicks/session (n)	**−1.01 ± 1.09**	**0.59 ± 0.82**	**0.99 ± 0.87**	**−0.38 ± 0.61**	**−0.12 ± 0.75**
Clicks/day (n)	**−1.51 ± 0.56**	**0.74 ± 0.63**	**−0.08 ± 0.68**	**−0.23 ± 0.50**	**1.33 ± 0.78**
Total time online (min)	**−1.48 ± 0.60**	**0.89 ± 0.72**	−0.21 ± 0.54	−0.18 ± 0.55	**1.24 ± 0.95**
Aver. duration of online sessions (min)	**−1.04 ± 1.16**	0.56 ± 1.01	0.56 ± 0.99	**−0.19 ± 0.73**	**0.17 ± 0.79**
Perspective 3: Procrastination					
Largest period of inactivity (h)	**0.31 ± 1.84**	**0.05 ± 0.74**	**0.67 ± 1.00**	**−0.12 ± 0.92**	**−1.02 ± 0.81**
Days with 0 clicks (% of period)	**0.48 ± 1.16**	**−0.15 ± 0.58**	**0.79 ± 0.76**	**−0.04 ± 0.79**	**−1.36 ± 0.68**
PercCourse_1Login	**0.55 ± 1.27**	−0.10 ± 0.68	**0.34 ± 0.91**	−0.06 ± **0.89**	**−0.68 ± 1.07**

As for the remaining perspectives, most of the results are consistent with the observations for undergraduate students for most strategies. Students with the highest level of activity (Strategy 5) presented the highest values for the *Time-on-task* perspective and the lowest for the *Procrastination* perspective. Likewise, the least engaged students (Strategy 1) consistently had the lowest values concerning *Time-on-task* and relatively high values in features in *Procrastination*. In learning Strategy 3, students adopting it were characterized by high levels in *Procrastination*, having the longest periods of inactivity and the greatest number of days without any activity. Although these students accessed Moodle infrequently, when they did, they tended to have long and click-intensive sessions. Despite having long sessions, they had a low number of sessions overall and spent less total time on Moodle.

Again, to facilitate interpretation, the strategies were labeled based on these engagement characteristics in a manner similar to the labels attributed to the undergraduate students. Strategy 1 was labelled *inactive*, Strategy 2 *mildly frequent and task-focused*, Strategy 3 *likely procrastinators*, Strategy 4 *mildly frequent, mild activity in short sessions* and Strategy 5 *active, prolonged and frequent*.

3.2 End-of-course performance for undergraduate and graduate students

The main focus of this second section was the exploration of the relationship between various learning strategies and student performance. A Welch's t-test was employed to compare the average end-of-course performance of each learning strategy against all others within the same level of instruction (Table 6). The analysis revealed significant differences in performance among the learning strategies identified by *k*-means clustering.

Specifically, three out of the five strategies showed a significant difference from all others in undergraduate enrollments. Strategies A (characterized by moderate frequency and activity in short sessions) and C (moderate frequency and task-focused) were not significantly distinct from each other, but they were significantly different from all other strategies (p-value = 0.14).

Table 6
*Pairwise comparison of the statistics and p-values obtained for the Welch's t-tests comparing the end of course grades obtained by each learning strategy (cells with p-value < 0.05 identified with *)*

Undergraduate learning strategies								
	Strategy A		Strategy B		Strategy C		Strategy D	
	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value
Strategy A (n = 883)								
Strategy B (n = 665)	6.50	1.21^{-10}*						
Strategy C (n = 735)	1.46	0.14	−4.65	3.75e^{-6}*				
Strategy D (n = 702)	−2.41	0.02*	−8.17	7.96e^{-16}*	−3.48	5.11e^{-4}*		
Strategy E (n = 545)	7.45	2.60e^{-13}*	2.50	0.01*	6.11	1.44e^{-9}*	8.71	1.68e^{-17}*
Graduate learning strategies								
	Strategy 1		Strategy 2		Strategy 3		Strategy 4	
	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value
Strategy 1 (n = 1381)								
Strategy 2 (n = 1697)	−4.36	1.34e^{-5}*						
Strategy 3 (n = 1503)	0.18	0.86	5.34	1.02e^{-5}*				
Strategy 4 (n = 2093)	−9.07	2.49e^{-19}*	−6.10	1.17e^{-09}*	−11.00	1.26e^{-27}*		
Strategy 5 (n = 1143)	−10.68	4.29e^{-36}*	−8.26	2.26e^{-26}*	−12.46	1.16e^{-34}*	−3.22	1.29e-^{34}*

A closer look at the performance of students who adopted each strategy (Figure 2) provides more insights. Students who adopted Strategy D (*active, prolonged, and frequent engagement*) achieved the highest average grade of 14.85 (± 3.29). They were closely followed by students employing Strategies A and C, with average grades of 14.46 (± 3.26) and 14.19 (± 3.94), respectively. On the other hand, students using Strategy B (*likely procrastinators*) had the second-lowest average grades (13.17 ± 4.09). Notably, students who adopted Strategy E (*inactive*), despite some exceptions indicated by the high standard deviation, generally achieved lower grades (12.41 ± 5.85) than their peers using other strategies.

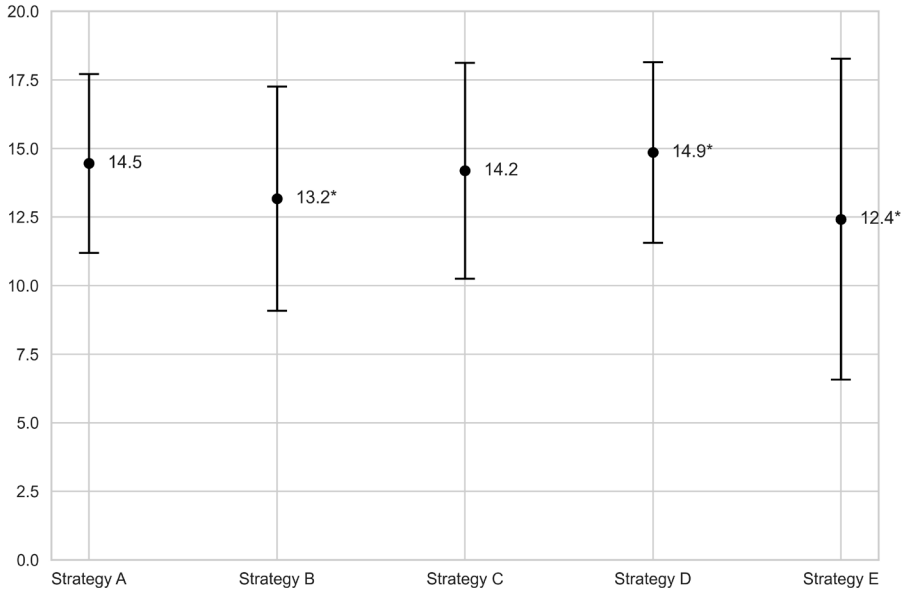

Figure 2

Average and standard deviation of the end-of-course grade for undergraduate learning strategies (values that exhibit statistically significant differences against all other groups are identified with an asterisk)

In the case of graduate students, three of the learning strategies were found to be significantly different from all others, with strategies 1 (*inactive*) and 3 (*likely procrastinators)* not showing significant distinction from each other (p-value = 0.86) while being significantly different from the remaining strategies. Figure 3 displays the average and standard deviation of the end-of-course grades for the graduate learning strategies identified by the *k*-means algorithm. Students who adopted learning Strategy 5 (*active, prolonged, and frequent engagement*) achieved the highest average grades (16.33 ± 2.74), followed by those adopting learning Strategy 4 (*mildly frequent, mild activity in short sessions*) with an average grade of 16.01 (± 2.77) and Strategy 2 (*mildly frequent and task-focused*) with an average grade of 15.46 (± 2.75). Strategies 1 and 3 were associated with the lowest average grades among all learning strategies used by graduate students, with average grades of 14.92 (± 3.85) and 14.90 (± 3.15), respectively.

This section has provided a detailed analysis of the relationship between various learning strategies and student performance. Significant differences in performance among the learning strategies were observed at both undergraduate and graduate levels. The data suggests that the choice of learning strategy can significantly impact academic performance.

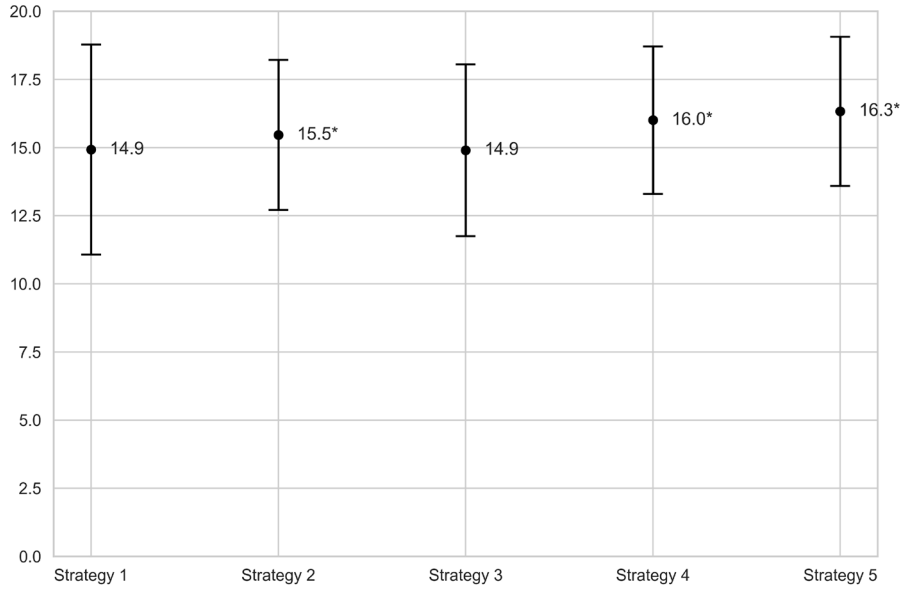

Figure 3
Average and standard deviation of the end-of-course grade for graduate learning strategies (values that exhibit statistically significant differences against all other groups are identified with an asterisk)

4 Discussion

4.1 Research question 1: What course-agnostic learner strategy profiles can be extracted from undergraduate and graduate students' SRL features extracted from LMS data?

The first research question in this study aimed to uncover course-agnostic learning strategy profiles from undergraduate and graduate students based on SRL features extracted from LMS data. The analysis identified five distinct profiles at each instruction level with varying levels of engagement, activity, and procrastination tendencies. The strategies identified were relatively similar for both the graduate and undergraduate levels.

The first learning strategy, *active, prolonged and frequent*, refers to students who were generally the most engaged across all perspectives. This learning strategy suggests that these students consistently devote time and effort to accessing the LMS and the materials contained therein, thus suggesting well developed SRL resource management skills (Pintrich et al., 1991). More specifically, regular and prolonged accesses hint at the students' awareness of the materials available and their ability to schedule the necessary time to

study (*Time and study environment*). Moreover, the frequent accessing also suggests discipline to continue studying over the entire semester, suggesting elevated *effort regulation*.

The second strategy, *mildly frequent, mild activity in short sessions*, is associated with students who logged into the LMS somewhat regularly but had short sessions with average levels of activity. The regular accesses also point to a certain degree of development in skills associated with *effort regulation* and *time and study environment*. While additional data would be needed to confirm this, the behavior exhibited by these students suggests that their main focus would be having the discipline to access specific materials deemed relevant, and logging out of the LMS afterwards, suggesting the existence of a more strategic approach, which was something observed in Cerezo et al.'s (2016) *Task-oriented and socially focused group*.

Students adopting the third learning strategy, *mildly frequent and task-focused*, showed average values for most activity metrics but specifically concentrated their efforts on completing assessments. This group shares certain similarities in learning strategy with the second group, with the main difference being the types of resources accessed by the students, which suggests some degree of development in skills associated with *effort regulation* and *time and study environment*. However, due to the partial nature of LMS data, it is impossible to draw meaningful distinctions between these two groups regarding SRL traits.

The fourth learning strategy, *likely procrastinators*, consisted of students who started interacting with course materials later, indicating procrastination. However, once logged in, they had long and intensive sessions, which aligns with conventional procrastination behavior, indicative of poor resource management skills, and has been shown to be a marker for poorer academic performance (Cerezo et al., 2016; Riestra-González et al., 2021; Yang et al., 2020).

The fifth and final learning strategy, termed *inactive*, is associated with students who exhibited the lowest LMS activity and engagement levels across all metrics. These students may be facing challenges that prevent them from engaging with the course materials or rely on resources outside of the LMS for their learning. Future research could focus on identifying the reasons behind such low engagement levels, in order to provide appropriate support and resources to better understand and address their needs.

Considering the results and the information presented in Table 1, it is possible to see differences in how students at the undergraduate and graduate levels behave on Moodle in absolute terms. However, in relative terms, the learning strategies they followed share similarities that do not warrant a meaningful distinction in their description. Thus, Research Question 1 can be answered by stating that *k*-means uncovered five distinct patterns

of interaction for learning strategies that were similar for both instruction levels: *active, prolonged and frequent engagement; mildly frequent and task-focused engagement*; *mildly frequent, mild activity in short sessions engagement*; *procrastinators;* and *inactive.*

4.2 Research question 2: What is the relationship between the learning strategies uncovered by k-means and end-of-course performance at each instruction level?

Baker et al. (2020) noted that clickstream data from LMS logs provide only a noisy and partial view of student behavior and learning. However, when the average end-of-course performance of students was mapped to their Moodle learning strategies, similar patterns were found for both undergraduate and graduate instruction levels.

Students who adopted the *inactive* learning strategy achieved the lowest grades, with an average of 12.41 for undergraduates and 14.90 for graduates. They were followed by those who adopted the *likely procrastinators* approach, with an average of 13.17 for undergraduates and 14.92 for graduates. These grades, in conjunction with the observed behavior on the LMS, suggest that some students in these groups either lacked a learning strategy with Moodle or had an inefficient approach to learning, both indicative of poor resource management skills development. These findings are consistent with other studies that have found lower levels of engagement to be associated with lower academic achievement (Cerezo et al., 2016; Hung & Zhang, 2008; Riestra-González et al., 2021; Yang et al., 2020). However, it is important to interpret these results with caution, as some students who did not engage with Moodle still obtained remarkable grades, possibly due to having a learning strategy that did not include active engagement with the LMS.

On the other hand, students who followed the *active, prolonged and frequent engagement* strategy achieved the highest overall grades. They were followed by those who adopted the *mildly frequent, mild activity in short sessions engagement* strategy, and those who followed the *mildly frequent and task-focused engagement* strategy. The evidence suggests that starting early and logging in frequently is an important factor in achieving better outcomes than the other strategies discussed previously. Although additional data would be needed for a more comprehensive assessment of these students, the behavior exhibited at least hints at the existence of a baseline learning strategy in place for the students' interactions with Moodle. An additional factor that may differentiate between grades are the types of actions performed on Moodle and the time spent on it. While it is true that the most successful students were also the most active, there is evidence that the types of interaction, rather than total activity, also play a relevant role in determining academic success. The results show that the two most successful strategies focused more on consulting theoretical content such as resources or external URLs. This is particularly interesting

because other studies (Cerezo et al., 2016; Riestra-González et al., 2021) found that students with a theoretical focus were surpassed by those who were equally engaged but followed a task-oriented approach, which was not the case for the present data. It is also important to note that not all time spent studying is equal, as noted by Cerezo et al. (2016). The second-most successful students clicked less and spent considerably less time on Moodle than their peers following the first and third-most successful approaches. This suggests that these students may have adopted a more strategic approach to their learning, resulting in a more efficient and higher quality use of their study time.

The findings from this study provide an answer to Research Question 2: A generally positive relationship was observed between the levels of engagement in learning strategies, as uncovered by *k*-means, and end-of-course performance across both instruction levels. Students who adopted *inactive* or *likely procrastinator* approaches to learning tended to have the lowest grades, while those who engaged in *active, prolonged, and frequent* interactions with Moodle achieved the highest overall grades. Early and frequent access to Moodle emerged as a key factor in achieving better outcomes. However, while this relationship was clear at the extreme ends of the spectrum, it became less distinct in the middle. Here, other factors such as the types of actions performed on Moodle and the time spent on it began to influence academic success in ways that were not always immediately apparent. Moreover, it is crucial to remember that Moodle logs represent only a portion of the learning process. This approach does not measure other potentially impactful factors, such as intrinsic motivation. Therefore, while Moodle logs provide valuable insights, they should be viewed in a broader context when evaluating student learning strategies and academic performance.

4.3 Research question 3: Are there differences in the effectiveness of the learning strategies between instruction levels?

When examining the clustering analysis results, there appear to be only minor differences between the learning strategies adopted by undergraduate and graduate students, as the same five general strategies emerged at both instruction levels. The primary difference was that, despite starting to access Moodle much later, undergraduate students exhibited higher overall levels of engagement in comparison to their graduate counterparts. From Table 1, we know that, on average, undergraduate students had higher amounts of clicks, sessions, and time spent on Moodle. These findings are also supported by the differences in prevalence of the different strategies at both levels. Approximately 25.01% of the undergraduate students adopted the *mildly frequent and task-focused engagement* strategy (against 21.71% in graduate students), while the most common learning strategy among graduate students is the

mildly frequent, mild activity in short sessions (26.78% compared to 20.82% of undergraduates). Graduate students also have a lower prevalence of *active, prolonged and frequent engagement* than their undergraduate counterparts (14.62% to 19.89%). These results align with expectations, as graduate students are generally older and are expected to have more developed resource management SRL skills, thus being more likely to efficiently manage their time and resources, and not needing to spend as much time logged in to fulfil their study objectives.

However, when examining the relative effectiveness of strategies at each instruction level, the patterns were remarkably similar. Across both groups, the ranking of learning strategies relative to their end-of-course grades followed the same order, with the strategies involving the most frequent accesses leading to the highest grades and procrastination and inactivity being associated with the lowest student performance. The consistency of these findings suggests that the core relationships between LMS engagement patterns and course outcomes are potentially generalizable across undergraduate and graduate contexts. While undergraduate students may utilize online platforms more extensively overall, the basic connections between behavior and performance appear to hold steady at both instruction levels.

Therefore, the answer to Research Question 3 is that no major differences were observed in the relative effectiveness of learning strategies between instruction levels. The key factors leading to positive outcomes remained important for both undergraduates and graduate students.

4.4 Implications

The findings presented herein provide relevant implications for both research and practice. On the research front, this work contributes to a growing body of literature aimed at uncovering learning strategies from trace data through unsupervised machine learning techniques. The results showcase both the potential and limitations of using LMS logs to categorize students based on their engagement patterns. In particular, the consistency of the relationships between strategy and performance across undergraduate and graduate contexts points to opportunities for developing more generalized models. Exploring the reasons behind students' choice of strategies is another area for future work, as the motivations and challenges faced by different learners, especially the less active ones, are still unclear. Qualitative or survey data collected alongside the logs may reveal additional insights into which motivational and emotional factors contribute to the understanding of some of the performance differences between strategies.

In practice, categorizing students into strategy profiles could inform the design of personalized interventions to improve resource management skills. Students following less successful approaches could receive prompts or

tutorials for developing better time management habits or content pacing. These adaptive supports would not be a one-size-fits-all solution; they would target the specific gaps exhibited through the engagement patterns. Moreover, course designers could use this knowledge to design programs and courses to promote forms of engagement that are more conducive to developing SRL skills and, more importantly, student success. Additionally, the presented methodology for extracting and analyzing variables from LMS data could be packaged into a reusable toolkit for institutions with accessible analytics dashboards that automatically cluster students based on trace behaviors, providing educators with actionable insights to refine their instructional practices and better support learners.

4.5 Limitations

This study has several limitations that must be acknowledged. The data source consists exclusively of LMS logs from a single institution over one academic year. While the sample size is large, incorporating multiple schools over longer periods could improve generalizability. Reliance on a unique data source also provides an incomplete picture of the learning process, as offline behaviors and other contextual variables are unavailable. Future research on this topic could complement data from the LMS with other instruments to develop a more comprehensive understanding of the learning strategy profiles.

Another relevant limitation concerns the SRL theoretical grounding of this approach. While theoretical connections are drawn between strategies, features, and SRL skills, all of them are indirect measurements of engagement with a single platform, and no direct observations of SRL constructs were performed. These connections, while suggested by empirical relationships, are not definitively confirmed. Future studies could incorporate established SRL instruments, such as the MSLQ, or use open-ended surveys or interviews. This could reveal individual motivations, challenges, and decision-making processes, providing a richer explanation for observed engagement patterns and performance differences. Such an approach could strengthen the theoretical basis of the analysis and offer nuanced insights into how students' SRL processes manifest in their online behaviors. Moreover, it could guide the development of interventions that target specific phases of the SRL process, thereby offering more targeted and effective support for students.

Conclusion

This work presented an analysis of uncovering learning strategies from Moodle log data through an unsupervised machine learning approach to assess learning strategy effectiveness across undergraduate and graduate contexts.

Clustering algorithms were leveraged to categorize over 11,000 student enrollments into distinct profiles based on their LMS engagement patterns. The findings revealed five similar strategies at both instruction levels: *active, prolonged and frequent engagement; mildly frequent and task-focused engagement*; *mildly frequent, mild activity in short sessions engagement*; *likely procrastinators*; and *inactive*.

Clear relationships emerged between engagement behaviors and student outcomes by mapping academic performance to these strategies. Across contexts, prolonged activity and early access were reliable markers of success, while procrastination and disengagement corresponded to lower achievement. However, success factors were more complex for some groups, involving strategic use of time and choice of activities. Still, the core patterns translating engagement to performance were strikingly consistent between undergraduates and graduates.

Nonetheless, this research makes valuable contributions. It demonstrates the feasibility of extracting meaningful learning strategy profiles from LMS data at scale across courses and instruction levels. The findings illustrate connections between online behaviors and performance. The findings also inform design principles for personalized interventions that target the development of successful learning strategies.

However, some limitations should be acknowledged. The study relied solely on clickstream data, providing an incomplete view of learning processes. Additional data on student demographics, prior achievement, and psychological factors like motivation could enrich the analysis. Adding this data would allow for a more comprehensive incorporation of the results presented herein into one of the existing SRL models (Panadero, 2017), which would not only provide a clearer interpretation of the results but would also contribute to an increased understanding of the motivational and emotional processes that lead students to adopt specific learning strategies. Moreover, the specific courses, instructors, and institutional contexts likely influenced the results. The sample was collected from an information management school, and replicating this approach across more diverse settings would strengthen conclusions about the potential generalizability of a course-agnostic approach.

There are several promising avenues for future work building on this research. One direction involves applying similar techniques to datasets across multiple institutions over longer timeframes. This could evaluate the consistency of findings and further establish generalizability of the relationships between online behaviors, strategy profiles, and achievement. Additionally, incorporating supplementary data sources beyond Moodle logs, whether institutional datasets or direct SRL measurements, holds potential for constructing more comprehensive learner models. Methodologically, exploring alternatives beyond k-means clustering, and developing personalized feedback

mechanisms tailored to strategy profiles may unlock new possibilities. These next steps emphasize the importance of understanding the factors influencing learning strategies and academic performance and, hopefully, translate analytics into positive pedagogical impact through interventions that develop effective self-regulated learning strategies among students.

In conclusion, this work contributes both methodologically and empirically to the growing body of literature on mining learner strategies from trace data. The findings provide a foundation for personalized interventions while highlighting opportunities for future research. Supplementing logs with additional data sources and perspectives would lead to more robust, generalizable, and actionable models.

References

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, *11*(24), Article 7238. https://doi.org/10.3390/su11247238

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, *17*(1), Article 13. https://doi.org/10.1186/s41239-020-00187-1

Bellur, S., Nowak, K. L., & Hull, K. S. (2015). Make it our time: In class multitaskers have lower academic performance. *Computers in Human Behavior*, *53*, 63–70. https://doi.org/10.1016/j.chb.2015.06.027

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*, Article 103999. https://doi.org/10.1016/j.compedu.2020.103999

Biggs, J. B. (1987). *Student approaches to learning and studying. Study process questionnaire manual.* Australian Council for Educational Research.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*(2), 161–186. https://doi.org/10.1016/S0959-4752(96)00015-1

Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education*, *33*, 24–32. https://doi.org/10.1016/j.iheduc.2017.01.004

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A Systematic Review. *The Internet and Higher Education*, *27*, 1–13. http://dx.doi.org/10.1016/j.iheduc.2015.04.007

Çebi, A., & Güyer, T. (2020). Students' interaction patterns in different online learning activities and their relationship with motivation, self-regulated learning strategy and learning performance. *Education and Information Technologies*, *25*(5), 3975–3993. https://doi.org/10.1007/s10639-020-10151-1

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

Coates, H., James, R., & Baldwin, G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary Education and Management*, *11*(1), 19–36. https://doi.org/10.1007/s11233-004-3567-9

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Gašević, D., Jovanović, 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), 113–128. http://dx.doi.org/10.18608/jla.2017.42.10

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Hung, J.-L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *Journal of Online Learning and Teaching*, *4*(4), 426–437.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259.

Li, L.-Y., & Tsai, C.-C. (2017). Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance. *Computers & Education*, *114*, 286–297. https://doi.org/10.1016/j.compedu.2017.07.007

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, & J. Neyman (Eds.), *Berkeley Symposium on Mathematical Statistics and Probability* (s. 281–297). University of California Press. https://www.cs.cmu.edu/~bhiksha/courses/mlsp.fall2010/class14/macqueen.pdf

Matcha, W., Gašević, D., Jovanović, J., Uzir, N. A., Oliver, C. W., Murray, A., & Gasevic, D. (2020). Analytics of learning strategies: The association with the personality traits. *LAK '20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 151–160. https://doi.org/10.1145/3375462.3375534

McKinney, W. (2017). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd Ed.). O'Reilly Media.

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. *American Journal of Distance Education*, *34*(2), 137–156. https://doi.org/10.1080/08923647.2020.1696140

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. The Regents of The University of Michigan.

Puzziferro, M. (2008). Online technologies self-efficacy and self-regulated learning as predictors of final grade and satisfaction in college-level online courses. *American Journal of Distance Education*, *22*(2), 72–89. https://doi.org/10.1080/08923640802039024

Riestra-González, M., Paule-Ruíz, M. del P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, Article 104108. https://doi.org/10.1016/j.compedu.2020.104108

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135–146. https://doi.org/10.1002/cae.20456

Santos, R. M., & Henriques, R. (2023). Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs. *Computers and Education: Artificial Intelligence*, *5*, Article 100175. https://doi.org/10.1016/j.caeai.2023.100175

Susac, A., Bubic, A., Kaponja, J., Planinic, M., & Palmovic, M. (2014). Eye movements reveal students' strategies in simple equation solving. *International Journal of Science and Mathematics Education*, *12*(3), 555–577. https://doi.org/10.1007/s10763-014-9514-4

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, *27*(4), 551–572.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 531–566). Academic Press. https://doi.org/10.1016/B978-012109890-2/50045-7

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Yang, Y., Hooshyar, D., Pedaste, M., Wang, M., Huang, Y.-M., & Lim, H. (2020). Predicting course achievement of university students based on their procrastination behaviour on Moodle. *Soft Computing*, *24*(24), 18777–18793. https://doi.org/10.1007/s00500-020-05110-4

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

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

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299–315). Routledge.