Exploring Self-Regulated Learning in Virtual Environments: An Experimental Clustering-Based Approach

Juliete A. R Costa[1,2], Geycy D. O. Lima[1,3], Rafael D. Araújo[1], Fabiano A. Dorça[1]

[1]Faculdade de Computação (FACOM)
Universidade Federal de Uberlância (UFU), Uberlândia, MG - Brasil

[2]Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas
(IFSULDEMINAS), Carmo de Minas, MG - Brasil

[3]Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas
(IFSULDEMINAS), Inconfidentes, MG - Brasil

{juliete.costa,geycy.lima}@ifsuldeminas.edu.br
{rafael.araujo, fabianodor}@ufu.br

Abstract. *Following the COVID-19 pandemic, there was a substantial increase in the volume of educational data generated in online environments. In this context, this study investigates signs of self-regulated learning (SRL) in virtual environments by applying educational data mining techniques to analyze student behavior. Data were collected from Moodle logs of a technical course offered by a federal public educational institution and underwent a preprocessing phase. Clustering algorithms such as K-Means, HDBSCAN, and Agglomerative Clustering were then applied to identify behavior patterns related to SRL. Differing from previous studies that mainly focused on student profiling or general engagement-performance correlations, this research explores how behavioral patterns revealed by clustering are directly associated with SRL indicators, with the results showing that HDBSCAN and K-Means were more effective in forming meaningful groups. The analysis revealed that students who exhibited stronger indications of SRL tended to achieve better academic performance, demonstrating greater engagement with learning resources, which was reflected in higher grades. This study contributes to a more nuanced understanding of SRL dynamics in virtual environments and highlights the potential of educational data mining techniques in identifying relevant behaviors, offering valuable insights for the development of pedagogical practices that promote student autonomy.*

1. Introduction

Virtual Learning Environments (VLEs) are online systems designed to support educational activities across different academic levels and domains. Among the most widely adopted platforms globally is Moodle [Moodle 2024], which offers a variety of tools and features aimed at enhancing academic performance, increasing student motivation, and reducing dropout rates.

The relevance of VLEs became even more pronounced in 2020, when the COVID-19 pandemic triggered unprecedented changes in education systems worldwide [World Health Organization 2020]. As institutions were forced to rapidly transition from face-to-face to online learning modalities to maintain social distancing, VLEs played a

DOI: 10.5753/sbie.2025.12462

crucial role in ensuring the continuity of instruction. This large-scale and abrupt shift also led to the generation of vast volumes of educational data, creating unique opportunities to analyze student behavior in digital learning environments. Within this context, researchers began to pay attention to the concept of Self-Regulated Learning (SRL).

SRL is a key construct in educational psychology that refers to students' ability to plan, monitor, and evaluate their learning processes independently. It encompasses cognitive, metacognitive, motivational, and emotional dimensions [Panadero 2017], and research has shown that students who effectively self-regulate tend to achieve better academic outcomes [Zimmerman and Martinez-Pons 1986]. In online learning environments, where learners are expected to take greater responsibility for their educational progress, SRL becomes especially critical.

Building on this, several authors such as [Zimmerman 2000] and [Panadero 2017] have proposed cyclical models of SRL, composed of distinct but interconnected phases and subprocesses. Although terminologies may vary, most models converge on three main phases: (a) Preparation (or planning), (b) Performance, and (c) Evaluation. The preparation phase involves task analysis, goal setting, and planning; the performance phase covers task execution and monitoring; and the evaluation phase focuses on reflection and adaptation, aimed at continuous improvement in future learning activities.

As researchers seek to better understand and support SRL processes, the field of Educational Data Mining (EDM) has emerged as a valuable ally. EDM focuses on developing and applying methods to analyze data generated in educational settings [Costa et al. 2020]. These techniques are particularly useful for uncovering patterns in learning behaviors, supporting students in developing self-regulatory skills, and ultimately improving the overall effectiveness of educational systems [Cavalcanti et al. 2018].

The growing adoption of VLEs, intensified by the pandemic, not only facilitated digital instruction but also resulted in the accumulation of massive datasets that can be leveraged by EDM techniques [Ramos et al. 2020]. According to [Shaun et al. 2011], analyzing these data allows researchers to gain deep insights into how students learn, the contexts in which learning occurs, and the factors that influence educational outcomes.

Extracting meaningful insights from such data involves a multi-stage process: preprocessing, application of data mining techniques, and post-processing of results [Costa et al. 2020]. The preprocessing stage ensures that the data are properly formatted and cleaned for analysis. Subsequently, various mining techniques, such as classification, regression, association rules, clustering, sequential pattern mining, and text mining, can be employed. Finally, the results must be interpreted in light of educational goals, which requires both domain expertise and statistical validation to support decision-making.

Among the various EDM techniques, clustering methods, classified as unsupervised learning approaches, stand out for their ability to identify patterns in unlabeled data. These methods aim to group similar objects within a given context and include algorithms based on partitional, hierarchical, and density-based principles.

To explore SRL in a real-world educational context, this study analyzed log data from Moodle, collected in a post-secondary technical course offered by a public institution. The objective was to identify signs of SRL through students' interactions with the platform's learning resources. To this end, clustering algorithms were employed as part

of an EDM strategy to reveal behavioral patterns without the need for predefined labels.

Specifically, three clustering algorithms were applied: K-Means, a widely used partitional method; Agglomerative Clustering, which follows a hierarchical approach; and HDBSCAN, a density-based algorithm capable of handling noise and detecting clusters of varying densities. Through the identification of distinct behavioral profiles, the study aimed to examine how different levels of student engagement relate to SRL indicators and academic performance, differing from related works that primarily focused on student profiling or general engagement–performance correlations.

The underlying hypothesis is that certain engagement behavior on Moodle, such as the regular and varied use of learning tools, may highlight self-regulation tendencies and point to better academic outcomes. Based on this premise, the study posed the following research questions:

- RQ1 What types of events recorded in Moodle logs can be considered indicative of self-regulated learning behaviors among students?
- RQ2 Which of the applied clustering algorithms performs best in identifying groups with distinct learning behavior patterns?
- RQ3 How are the groups identified by the clustering algorithms related to students' final academic performance?

This paper is structured as follows: Section 2 presents the related works. Section 3 describes the materials and method used in this research, followed by the presentation and discussion of results in the section 4. Finally, Section 5 outlines the conclusions, limitations, and future works of the study.

2. Related works

Recently, several studies have explored SRL by analysing digital traces in VLEs. Among the different possible approaches, educational data mining techniques have stood out. In particular, unsupervised learning methods, such as clustering algorithms, are widely used to identify groups of students with different levels of SRL in online educational contexts [Damayanti et al. 2023].

The work by [Davies et al. 2021], for example, investigates the learning strategies adopted by students in an online course, using the k-means algorithm to categorise behaviours over time. The results show that some students adapt their strategies as the course progresses, which is evidence of self-regulation processes.

The study described in [Ramos et al. 2020] uses quantified Moodle records to identify student profiles in a distance learning course. Hierarchical and non-hierarchical clustering techniques were used to segment the data and identify three distinct student profiles: low, medium, and high interaction with the educational environment. Similarly, [Farida and Sudibyo 2022] examines the relationship between self-regulation and academic performance, grouping students via k-means. The research identifies three groups (low, medium, and high SRL) and a positive correlation between higher levels of SRL and better grades.

In [Nuankaew et al. 2022]'s study, SRL styles are investigated in a hybrid context, using the k-means, k-medoids, and x-means algorithms to group students based on

their behaviours. The data includes interactions on the platform, diagnostic tests (pre- and post-tests), and the results also reveal positive correlations between SRL and performance. [Peraić and Grubišić 2023] analyses patterns of engagement in an Introduction to Programming course, based on Moodle logs collected over three years. Using k-means, two groups were identified: students with high engagement and good performance, and those with low engagement and poor performance.

Finally, [Rodriguez et al. 2021] studies SRL based on video clicks and time management, identifying four distinct patterns using the k-means algorithm. The results show that planning activities in the early stages is related to better grades.

Unlike previous studies that primarily focus on identifying student profiles or general engagement-performance correlations, this paper investigates how the behavioural patterns revealed by clustering algorithms are directly related to indicators of self-regulated learning. By analysing student performance within each group, the study offers a detailed view of how SRL manifests in academic outcomes, providing evidence of self-regulation based on VLE interaction data.

3. Materials and Methods

The method adopted in this study has an experimental nature and is summarized in Figure 1, comprising four distinct stages. In the first stage, user interaction data (logs) from the Moodle platform were extracted, resulting in three main files: (1) the Grade Report, which includes students' scores for each activity and their final course grade; (2) the Log Report, which records all events performed by users in the course, including usage of available resources; and (3) the Configurable Reports plugin report[1], which details the amount of time each user dedicated to each course.

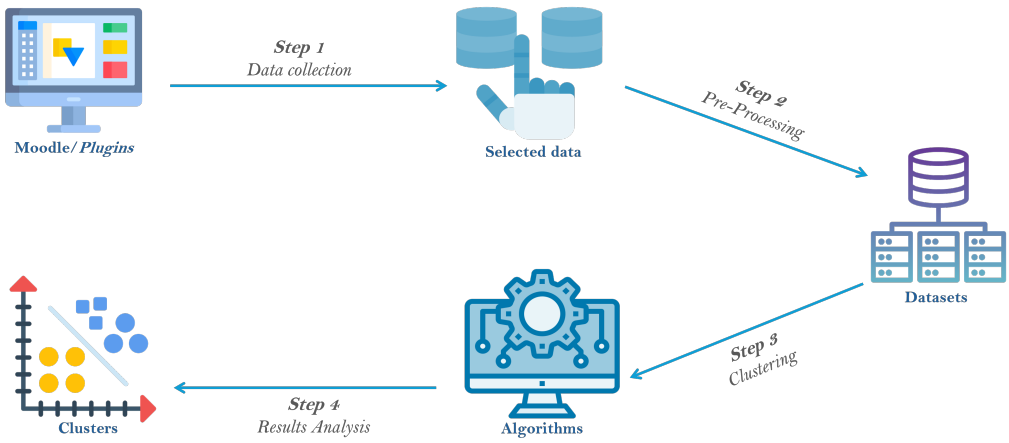

Figure 1. Stages of the method adopted in this study

In the next step, the data were preprocessed in order to select the relevant attributes and build the dataset to be used in the subsequent analysis. Then, clustering algorithms — specifically Agglomerative Clustering, K-Means, and HDBSCAN — were applied using appropriate libraries. Finally, the results were analyzed to identify signs of self-regulated learning (SRL) among students, enabling a deeper understanding of interaction and performance patterns in the virtual learning environment.

[1]https://moodle.org/plugins/block_configurable_reports

3.1. Step 1: Data collection

The data collected from Moodle referred to eight subjects taken in the first semester of the Online Technical Degree in Business Administration, a three-semester program offered by a federal public educational institution[2]. For each subject, we obtained three files in CSV format that were combined using the unique identifier of each student. The result was a consolidated file containing information on all active students in the subjects, including the quantification of all *logs* events and the time of access to the platform for each user. Table 1 presents a description of the data found, including the subjects, the number of *logs* generated, the number of filtered logs, the number of students, and the number of events generated who attended each subject.

Table 1. Subjects Description

Tag	Subject	Total Logs	Total Filtered Logs	Students	Events Logs
COURSE_1	Customer Service and Consumer Rights	76229	62564	413	25
COURSE_2	Entrepreneurship	108236	88717	456	27
COURSE_3	IT and Spreadsheets	186703	120021	445	40
COURSE_4	Introduction to Administration	78802	64521	409	29
COURSE_5	Labour and Social Legislation	97410	79869	418	27
COURSE_6	Business Model Canvas	78176	64518	405	35
COURSE_7	Sustainable Business	74255	60972	413	25
COURSE_8	Recruitment and Selection	136997	111919	430	25

3.2. Step 2: Pre-processing

The Pandas library in Python was used for data preprocessing and dataset construction. This library provides flexible data structures and robust analytical tools, enabling the efficient handling and manipulation of large volumes of data [McKinney et al. 2010].

Moodle records a wide variety of attributes in its activity logs, the quantity and types of which may vary depending on the platform's specific configurations and any additional plugins installed. In the Moodle environment used in this study, 16 attributes were identified as consistently present across all subjects, as shown in Table 2. It is important to note that attributes 1, 2, 3, and 16 do not directly correspond to activity log entries within the system. Nevertheless, for the construction of the dataset, both these attributes and the remaining common ones, along with the specific features of the resources available in each subject, were initially considered.

Subsequently, the files were then analyzed to identify distinct events for each subject and their impact on the dataset. Each subject offered in the program features specific configurations of activities and resources. Therefore, in addition to the common attributes listed in Table 2, additional attributes were identified. Supplementary attributes related to differentiated activities proposed by the instructor were included, as they represented alternative assessments distinct from quizzes.

[2]Data collection was approved by the Ethics Committee for Research with Human Beings of IFSUL-DEMINAS, under approval number CAAE 78890524.5.0000.8158

Table 2. Description of common attributes

Attribute	Type	Description
[1]:id	Numeric	The student's register number.
[2]:name	Text	The student's name
[3]:grade	Numeric	Attribute with the student's final grade.
[4]:some_content_published	Numeric	Represents how many times the student has published content in activities within the course.
[5]:post_created	Numeric	Indicates the number of posts the student has created in specific activities, such as forums, blogs, or other discussion areas.
[6]:curso_viewed	Numeric	Indicates the number of times the student has viewed the course.
[7]:discussion_viewed	Numeric	Indicates the number of times the student has viewed discussions, such as forums or discussion groups.
[8]:module_course_viewed	Numeric	Shows how many times the student has viewed a specific module within the course.
[9]:summary_attempt_ questionnaire_viewed	Numeric	Indicates the number of times the student viewed the summary of quiz attempts.
[10]:attempt_questionnaire _viewed	Numeric	Indicates how many times the student viewed quiz attempts.
[11]:completion_activity _course	Numeric	Represents how many course activities the student has completed.
[12]:report_of_grades_viewed	Numeric	Indicates how many times the student has viewed the grade report for the course.
[13]:attempt_questionnaire _delivered	Numeric	Indicates how many quiz attempts the student has submitted.
[14]:attempt_questionnaire _started	Numeric	Indicates how many quiz attempts the student has started.
[15]:attempt_questionnaire _reviwed	Numeric	Indicates the number of times the student has revised quiz attempts.
[16]:time	Numeric	Represents the total duration (in seconds) the student spent on the subject.

To examine the relationship between variables, a correlation matrix was constructed using Spearman's non-parametric coefficient. This method was selected because the data did not follow a normal distribution, as previously determined by the Kolmogorov–Smirnov normality test, which yielded $p\_value < 0.05$ across all subjects analyzed.

The Spearman coefficient was used to measure monotonic relationships between variables [De Winter et al. 2016, Spearman 1961]. Correlation analysis was applied to all datasets, and attributes with strong correlations (> 0.7) were aggregated accordingly [Zar 2005]. Attributes with low system usage or many missing values were excluded. After preprocessing, five courses resulted in datasets with five attributes, and three courses had six, due to variations in recorded logs. These refined datasets enabled the next phase: applying clustering techniques to identify student behavior patterns.

Table 3 summarizes the final attributes, organized according to the phases of the SRL model [Zimmerman 2000, Panadero 2017]. In the Preparation phase, involving goal

setting and task analysis, we associated the *views* attribute with preparatory behavior, as it reflects students' efforts in accessing materials early on. The *posts*, *quizzes*, and *submission* attributes were linked to the Performance phase, capturing active engagement during task execution. We also included *completed_activities* here, as it records real-time task completion. For the Evaluation phase, which emphasizes reflection and adaptation, we associated the *time* attribute as an indirect indicator of student effort and persistence, since time-on-task can suggest ongoing engagement even in the absence of observable reflective actions. Additionally, *completed_activities* supports this phase when seen as a cumulative indicator of progress, aiding self-assessment and strategic adjustments.

Table 3. Attribute Description and SRL Phase Association

Attributes	Description	SRL phase
id	User identifier attribute.	Not applicable
posts	Total student publications throughout the course, including forum posts and activity comments.	Performance
views	Total student views of course resources, including modules, quizzes, materials, activities, comments, and grades.	Preparation
completed_activities	Number of course activities completed and/or updated.	Performance/ Evaluation
quizzes	Total quiz-related actions by the student, including starting, reviewing, and submitting quizzes.	Performance
submission	Total number of evaluative submissions by the student, including files or online texts to the teacher. Attribute only present in courses 3, 4 and 6.	Performance
time	Time in seconds that the student dedicated to the course using the platform.	Evaluation

3.3. Step 3: Clustering

This study examines the effectiveness of three clustering algorithms, K-Means, Agglomerative, and HDBSCAN, to identify indicators of students' self-regulated learning from Moodle logs. K-Means and Agglomerative were chosen for being the most commonly used [Aldowah et al. 2019, Salloum et al. 2020], while HDBSCAN was included as an alternative to the density-based DBSCAN, which showed low performance on the analyzed data.

The K-Means algorithm was selected for being an efficient partitional method that segments data into groups based on the mean of the data points, facilitating the identification of well-defined clusters. Agglomerative Clustering, in turn, is a hierarchical approach that builds a clustering tree (dendrogram) based on data similarity, allowing for a more detailed analysis of the structural relationships among data points. HDBSCAN was included due to its ability to handle data with varying densities and to identify outliers, an essential feature for capturing the diversity of student behaviors in VLEs. The use of these algorithms enabled a robust and comprehensive analysis aimed at detecting groups of students with indicators of self-regulated learning, providing valuable insights into engagement patterns and academic performance.

3.4. Step 4: Results Analysis

For this step, we consider the Silhouette Coefficient, Calinski-Harabasz, and Davies-Bouldin metrics to evaluate the internal quality of the clusters. In the case of the HDBSCAN algorithm, the number of outliers identified was also taken into account. Results

from this algorithm, despite similar validation scores, were excluded for classifying too many records as outliers.

The Silhouette Coefficient is a metric that assesses how well each data point fits within its assigned cluster, taking into account both the distance to other points within the same cluster and the distance to points in different clusters [Dinh et al. 2019, Rousseeuw 1987]. This metric was selected because it provides a comprehensive evaluation of clustering quality, considering both internal cohesion and separation between neighboring clusters. The Calinski-Harabasz index measures group density and separation to identify compact clusters, while the Davies-Bouldin index assesses similarity to neighboring clusters, highlighting well-separated groupings [Furlanetto et al. 2022].

In addition to internal validation measures, statistical significance tests were applied to compare the clusters identified by the algorithms, aiming to determine whether there are significant differences among them. Finally, an analysis of self-regulation evidence within each cluster was conducted, and the relationship between these groups and students' final performance in each course was examined.

4. Results

The experimental results generated by the three clustering algorithms were analyzed to identify the most suitable algorithm and determine the optimal number of clusters. Table 4 shows, for each dataset, the clustering algorithm that achieved the best validation metrics and the corresponding results. The Agglomerative Clustering algorithm is not included in the table, as it did not achieve satisfactory internal validation scores in any of the datasets. The Figure 2, in turn, presents boxplots that highlight the differences in the mean values of each observed attribute across the clusters identified by the algorithms for each course. These visualizations illustrate how the attributes vary between clusters, emphasizing potential patterns or distinctions. To evaluate whether the observed mean differences were statistically significant, the Kruskal-Wallis test was applied.

The best result for *COURSE_1* was obtained by the HDBSCAN algorithm. Both HDBSCAN and K-Means achieved the same result for the Silhouette measure. Still, HDBSCAN was able to separate the data more effectively by identifying 66 records as outliers, leading to higher values in the C-H and D-B indices. Furthermore, the mean differences between the identified clusters were statistically significant across all attributes. In *COURSE_2*, HDBSCAN also stood out as the best option, showing superior validation metrics and statistically significant cluster separation across all attributes.

In *COURSE_3*, the K-Means and HDBSCAN algorithms produced similar clustering results, with K-Means showing superior performance across the three validation metrics. As illustrated in Figure 2, the boxplots reveal that the values for the *quizzes* attribute are quite similar between *cluster0* and *cluster1*, which aligns with the lack of statistical significance identified by the Kruskal-Wallis test. This outcome may be explained by the instructor's emphasis on alternative assessment methods, such as the submission of online texts and files, captured by the *submission* attribute, rather than quizzes, in this particular course. In *COURSE_4*, the K-Means and HDBSCAN algorithms had similar results for the Silhouette Coefficient, but HDBSCAN outperformed K-Means on the C-H and D-B indices. Additionally, the average values observed in *cluster1* were higher than those in *cluster0*, with statistical significance across all attributes.

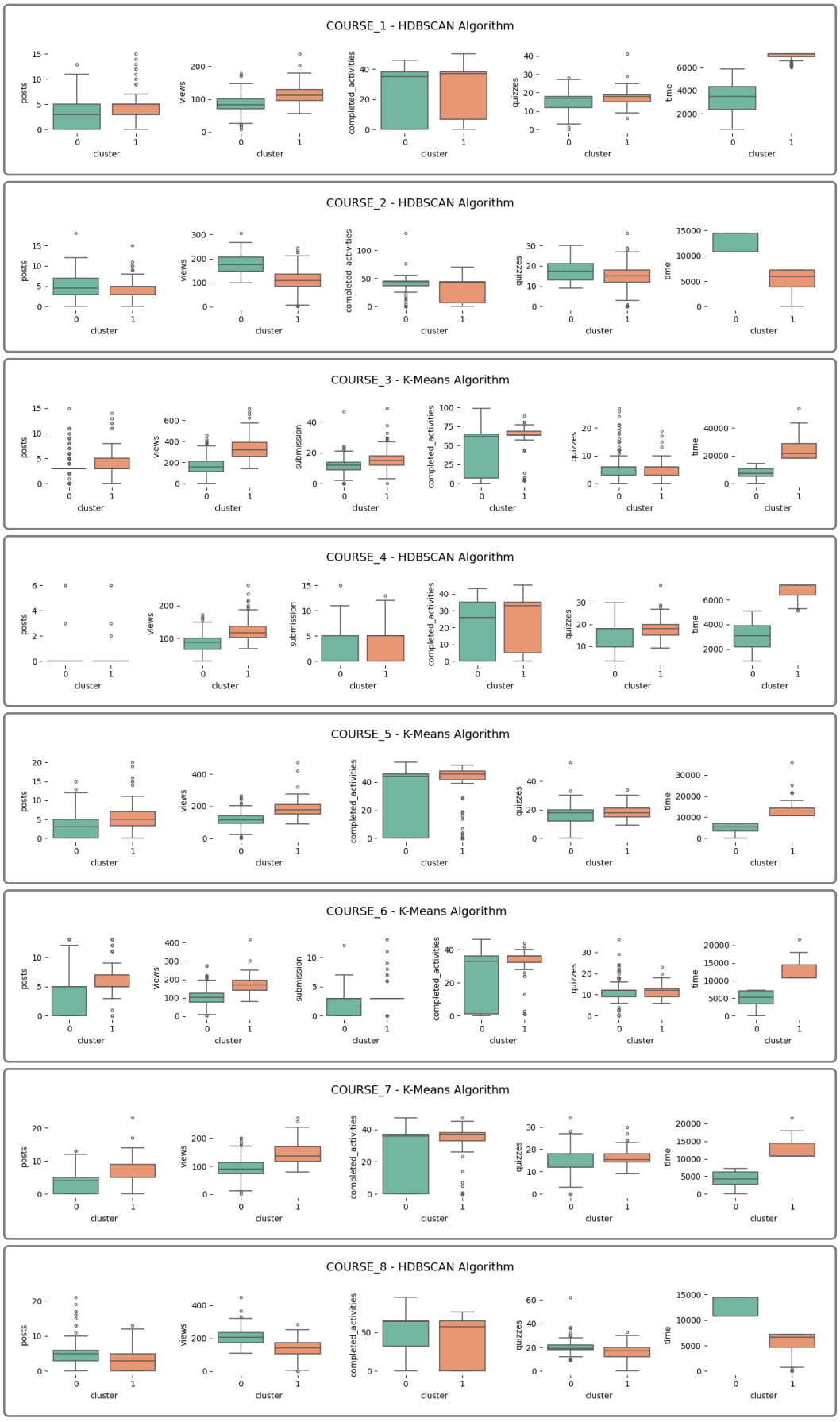

Figure 2. Boxplot analysis of the groups found by clustering algorithms

Table 4. Clustering validation metrics for different algorithms

Tag	Top-performing algorithm	Silhouette	C-H	D-B	Clusters
COURSE_1	HDBSCAN	0.656	892.35	0.378	cluster0(N=227) cluster1(N=120) Outliers(N=66)
COURSE_2	HDBSCAN	0.695	1043.02	0.495	cluster0(N=110) cluster1(N=317) Outliers(N=29)
COURSE_3	K-Means	0.685	946.39	0.494	cluster0(N=355) cluster1(N=90)
COURSE_4	HDBSCAN	0.691	1230.68	0.400	cluster0(N=179) cluster1(N=143) Outliers(N=87)
COURSE_5	K-Means	0.680	787.64	0.563	cluster0(N=308) cluster1(N=110)
COURSE_6	K-Means	0.685	742.29	0.521	cluster0(N=335) cluster1(N=70)
COURSE_7	K-Means	0.699	710.37	0.487	cluster0(N=361) cluster1(N=52)
COURSE_8	HDBSCAN	0.701	948.99	0.517	cluster0(N=131) cluster1(N=219) Outliers(N=80)

C-H: Calinski-Harabasz Index; D-B: Davies-Bouldin Index.

In courses *5*, *6*, and *7* showed better results using the K-Means. In the cases of courses *5* and *6*, the Silhouette scores were similar for both K-Means and HDBSCAN, whereas the C-H and D-B indices were slightly better for HDBSCAN. However, the presence of a significant number of outliers in HDBSCAN compromised the analysis. In *COURSE_7*, both algorithms showed similar separation and statistical significance, but K-Means had better Silhouette and C-H results.

In *COURSE_6*, the *quizzes* attribute once again did not present statistically significant differences between clusters, which may be related to the instructor's adoption of alternative assessment methods, evidenced by the presence of the *submission* attribute in the dataset. Similarly, in *COURSE_7*, no statistical significance was observed for the *quizzes* attribute. However, the remaining attributes exhibited higher values in *cluster1* compared to *cluster0*, despite the former comprising only 52 students. Finally, in *COURSE_8*, K-Means and HDBSCAN had similar separation and significance, but HDBSCAN outperformed K-Means on all three validation metrics. Additionally, *cluster1* showed higher average values than *cluster0* across all attributes with statistical significance.

We observed that, across all analyzed courses, the clusters with fewer students exhibited higher average values for all evaluated attributes. In each course, this cluster was classified as the group with indications of self-regulated learning, as it included students who demonstrated greater activity within the system throughout the course offering. Specifically, these students made more forum and/or activity posts, accessed course content and activities more frequently, submitted more assignments and quizzes, and consequently dedicated more study time during the course period.

Additionally, we found that in only four courses, the mean differences between groups did not reach statistical significance for at least one attribute, most commonly the

quizzes attribute. This result suggests that quiz participation tends to be uniform among students. However, the other attributes, number of views, posts, completed activities, and time dedicated to the course, were consistently higher across all datasets for the group with indications of self-regulated learning. These findings reinforce that, beyond completing quizzes, these students engage more actively with other system resources.

In each cluster, the percentage of students with Grade C (students who failed the course), Grade B (students who passed with an average between 6 and 8), and Grade A (students who passed with an average above 8) was analyzed. The panel presented in Figure 3 highlights the performance analysis of the more engaged groups, which also exhibit indications of self-regulated learning, within the clusters identified for each course. The Mann-Whitney U statistical significance test [Urdan 2010] was applied to assess whether the grades of students in the cluster identified as more engaged and self-regulated differed significantly from those in the other cluster generated by the algorithm. Before this analysis, it was verified that the grade distributions within each cluster did not follow a normal distribution, as confirmed by the Kolmogorov-Smirnov test.

The Mann-Whitney U test, a non-parametric method, assesses whether two independent samples differ significantly. It tests the Null Hypothesis (H_0) of no difference and the Alternative Hypothesis (H_1) of a difference. Grades were compared between each *cluster* per course, and in all cases, *p-values* were below 0.05, rejecting H_0 and indicating statistically significant differences between clusters.

Figure 3 illustrates that, across all the analyzed courses, the cluster composed of more engaged students, those showing indications of self-regulated learning, contains a higher percentage of students with Grade A and a lower percentage with Grade C. These findings suggest that students belonging to groups characterized by signs of self-regulated learning tend to achieve higher academic performance. The results of the Mann-Whitney U significance test support this outcome and provide an answer to research question QP3, emphasizing a relationship between the engagement and self-regulatory characteristics identified through clustering techniques and students' academic achievement.

5. Conclusion and Future Works

This study investigated the SRL of students enrolled in technical course at a public educational institution through the analysis of interaction data collected from the Moodle platform using EDM techniques. The data underwent a rigorous preprocessing stage and were analyzed using clustering algorithms: K-Means, HDBSCAN, and Agglomerative.

The main findings of this study include the analysis of event logs recorded in Moodle and their preprocessing, resulting in refined datasets that enabled the identification of student interaction and performance patterns. The clustering algorithms—K-Means, HDBSCAN, and Agglomerative Clustering—were compared, revealing a tie between K-Means and HDBSCAN, with each algorithm achieving better results in four datasets. It was observed that students belonging to clusters with indications of self-regulated learning generally achieved higher academic performance, as the clusters with greater engagement in learning resources had a higher proportion of students with top grades.

This study has some limitations. The main one is that the identification of SRL behaviors was based solely on log data, which may not capture all the nuances of students'

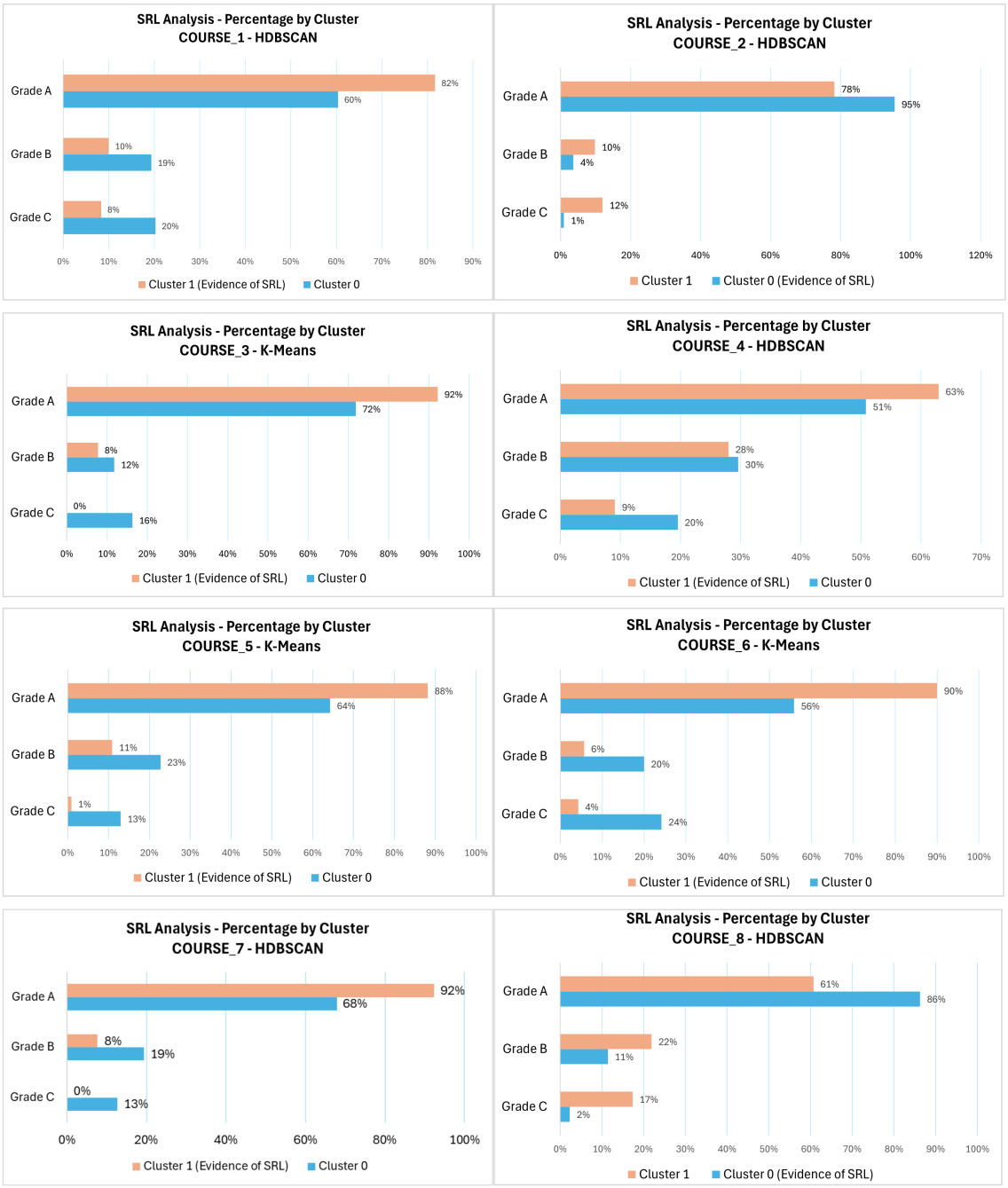

Figure 3. Analysis of SRL evidence in each cluster found by the algorithms

learning behavior. Furthermore, the analysis was limited to a specific set of courses and a single institutional context, which may constrain the generalizability of the results.

For future work, a more detailed analysis of HDBSCAN outliers is recommended, as the algorithm revealed noteworthy findings. Integrating other data sources, such as questionnaires and interviews, could complement log data for a more comprehensive view of SRL. These findings can also support the development of learning analytics tools, offering feedback and designing recommendation approaches that help students set goals, plan activities, and monitor progress.

Acknowledgment

The authors would like to thank the Federal University of Uberlândia (UFU) and the Federal Institute of Southern Minas Gerais (IFSULDEMINAS) for their support in the development of this research.

References

Aldowah, H., Al-Samarraie, H., and Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37:13–49.

Cavalcanti, A., Dourado, R., Rodrigues, R., Alves, N., Silva, J., and Ramos, J. L. C. (2018). An analysis of self-regulated learning behavioral diversity in different scenarios in distance learning courses. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1493.

Costa, J., Dorça, F., and Araújo, R. (2020). Avaliação do comportamento de estudantes em um ambiente educacional ubíquo. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 182–191, Porto Alegre, RS, Brasil. SBC.

Damayanti, A., Kusumawardani, S. S., and Wibirama, S. (2023). A review of learners' self-regulated learning behavior analysis using log-data traces. In *2023 IEEE 12th International Conference on Engineering Education (ICEED)*, pages 90–95. IEEE.

Davies, R., Allen, G., Albrecht, C., Bakir, N., and Ball, N. (2021). Using educational data mining to identify and analyze student learning strategies in an online flipped classroom. *Education Sciences*, 11(11):668.

De Winter, J. C., Gosling, S. D., and Potter, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273.

Dinh, D.-T., Fujinami, T., and Huynh, V.-N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *International Symposium on Knowledge and Systems Sciences*, pages 1–17. Springer.

Farida, A. and Sudibyo, N. A. (2022). Implementation of the k-means algorithm on learning outcomes and self-regulated learning. *UNION: Jurnal Ilmiah Pendidikan Matematika*, 10(2):147–154.

Furlanetto, G., Carvalho, V., Baldassin, A., and Manacero, A. (2022). Algoritmos de agrupamento aplicados à detecção de fraudes. In *Anais da XIII Escola Regional de Alto Desempenho de São Paulo*, pages 29–32, Porto Alegre, RS, Brasil. SBC.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Moodle (2024). Registered Moodle sites. Registered Moodle sites. Disponível em https://stats.moodle.org/sites/index.php?country=BR.

Nuankaew, P., Nasa-Ngium, P., and Nuankaew, W. S. (2022). Self-regulated learning styles in hybrid learning using educational data mining analysis. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*, pages 208–212. IEEE.

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology*, 8:422.

Peraić, I. and Grubišić, A. (2023). Exploring student engagement in online programming courses: A two-level k-means analysis. In *2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE.

Ramos, J., Santos, L., Silva, J., and Rodrigues, R. (2020). Identificação de perfis de interação de estudantes de educação a distância por meio de técnicas de agrupamentos. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 932–941, Porto Alegre, RS, Brasil. SBC.

Rodriguez, F., Lee, H. R., Rutherford, T., Fischer, C., Potma, E., and Warschauer, M. (2021). Using clickstream data mining techniques to understand and support first-generation college students in an online chemistry course. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 313–322.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Salloum, S. A., Alshurideh, M., Elnagar, A., and Shaalan, K. (2020). Mining in educational data: review and future directions. In *The International Conference on Artificial Intelligence and Computer Vision*, pages 92–102. Springer.

Shaun, R., Baker, J., Isotani, S., Maria, A., and Carvalho, J. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19:3–13.

Spearman, C. (1961). The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471.

Urdan, T. (2010). *Statistics in Plain English, Third Edition*. Taylor & Francis.

World Health Organization (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. World Health Organization. Available at https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020.

Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Zimmerman, B. and Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23:614–628.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation*, pages 13–39. Elsevier.