Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 18/08/2025; 1st round notif.: 24/03/2026; New version: 24/04/2026; 2nd round notif.: 11/05/2026;
Camera ready: 25/05/2026; Edition review: 28/05/2026; Available online: 28/05/2026; Published: 28/05/2026;

Self-Regulated Learning Traits in Students' Behavior Interactions in a Ubiquitous Learning Environment

Juliete Aparecida Ramos Costa
IFSULDEMINAS
ORCID: 0000-0002-2580-7869
juliete.costa@ifsuldeminas.edu.br

Geycy Dyany Oliveira Lima
IFNMG – Campus Salinas
ORCID: 0000-0002-3215-0367
geycy.lima@ifnmg.edu.br

Rafael Dias Araújo
Universidade Federal de Uberlândia
ORCID: 0000-0003-0545-2519
rafael.araujo@ufu.br

Fabiano Azevedo Dorça
Universidade Federal de Uberlândia
ORCID: 0000-0003-3281-0246
fabianodor@ufu.br

Abstract

Computers have become an integral part of everyday life. In education, these technologies enable the creation of Ubiquitous Learning Environments, which enrich the learning experience by providing more dynamic and engaging contexts, both in-person and online. Educators and institutions increasingly adopt technological tools to strengthen teaching and learning processes. As a result, the extensive use of these technologies has produced valuable data repositories that can be explored through data mining methods. In this context, this study presents an exploratory analysis of interaction data collected from a Ubiquitous Learning Environment, using Educational Data Mining techniques. Clustering methods were applied to explore students' behavior in learning sessions and to identify patterns associated with their quiz performance in the environment. Three data mining algorithms were applied to five distinct datasets, each prepared with different preprocessing strategies. The results showed that clustering performance is highly sensitive to data preprocessing, with the best outcomes achieved using feature selection and aggregation techniques. The findings showed statistically significant distinctions among the clusters and uncovered evidence of self-regulated learning within one of the groups. However, the effect size analysis indicated that only a subset of attributes, particularly those related to study time and interaction, presented substantial practical differences between clusters. These results suggest that students exhibiting higher levels of interaction and engagement tend to achieve better performance, indicating behavioral patterns potentially associated with self-regulated learning.

Keywords: *Ubiquitous Learning Environment; Educational Data Mining; Clustering; Self-Regulated Learning.*

Cite as: Costa, J. A. R., Lima, G. D. O., Araújo, R. D., & Dorça, F. A. (2026). Self-Regulated Learning Traits in Students' Behavior Interactions in a Ubiquitous Learning Environment. Revista Brasileira de Informática na Educação, vol. 34, pp. 770–798. https://doi.org/10.5753/rbie.2026.6368.

1 Introduction

Computational and technological tools have found extensive application across various fields of knowledge. These tools are becoming increasingly embedded in people's routines, helping them carry out everyday activities. This idea aligns with the notion of Ubiquitous Computing proposed by Weiser (1991). In the specific sphere of education, computers have been integrated into classrooms to create more interactive and stimulating learning environments, thereby supporting effective teaching and learning practices.

Such spaces, known as Ubiquitous Learning Environments (ULE), integrate resources that blend actions performed in both real and digital contexts, generating more meaningful study artifacts that account for students' contexts and individual traits (Zhao & Okamoto, 2011). Besides enabling these materials to be accessed from any place at any time, learning environments have evolved to become more intelligent, delivering customized and captivating educational resources that address the unique needs and differences of learners (Kinshuk et al., 2016). These advanced spaces are referred to as Smart Learning Environments (SLE) (Kinshuk et al., 2016).

Recent advances in Artificial Intelligence (AI) have further enhanced the capabilities of these environments, enabling more adaptive, data-driven, and personalized learning experiences. Nevertheless, recognizing and preserving each student's characteristics within such systems remains a complex task and an ongoing challenge (R. Baker et al., 2020). In this context, understanding how students regulate their own learning process becomes essential. Self-Regulated Learning (SRL) refers to the ability of learners to actively control their cognition, motivation, and behavior during the learning process (Panadero, 2017).

Some works have noted the importance of integrating technological mechanisms into learning environments to foster the self-regulated learning process (Viberg et al., 2020). In this process, the student is the protagonist of their learning and can develop several cognitive strategies, metacognition, motivation, and emotion/affection to self-regulate their learning (Panadero, 2017). However, accurately capturing and measuring SRL remains a significant challenge. R. Baker et al. (2020) points out that self-regulated learning is usually measured from students' self-reported data, and these self-reports alone may not be effective, since many individuals have insufficient or even biased memories.

This limitation reveals an important research gap: the lack of objective and fine-grained approaches to analyze SRL based on students' actual interactions within learning environments. In this sense, interaction data such as clickstream logs from VLEs provide a promising alternative, as they allow the observation of students' real behavior in a detailed and continuous manner.

In this sense, students' characteristics such as cognition, emotion, interests, and experiences are represented through a Student Model (Self, 1990), which is a fundamental component that provides subsidies to the computer system so it can personalize the content according to each student to make the learning experience more enjoyable and effective.

Among the various possible techniques in this context, AI methods, particularly those from the field of Educational Data Mining (EDM), can be used to understand how students learn and identify behavioral patterns that influence the learning process (R. Baker et al., 2020). Building on this, this work aims to analyze students' behavior during learning sessions based on their

interactions in a ULE, in order to better understand how they use the learning environment and its impact on their learning process.

The main contributions of this study are: (i) the investigation of different data transformation techniques for modeling learning sessions based on clickstream data; (ii) the comparison of multiple clustering algorithms to identify behavioral patterns; and (iii) the analysis of how these patterns relate to indicators of self-regulated learning.

This study is an extension of previous work (Costa et al., 2020), in which we analyzed students' clickstream data in a ubiquitous educational environment to understand their behavior during platform access sessions. In that prior work, we applied only the K-Means clustering algorithm and considered a single data transformation. In the current study, we expand this approach by incorporating analyses with five distinct data transformations and by adding two more clustering algorithms: agglomerative hierarchical clustering and the density-based HDBSCAN. The click records of students from a public university were organized into learning sessions, and the most frequently used resources were identified to shed light on the learning dynamics in this context.

To conduct this study, we raised the following research questions:

- RQ1: Which data transformation techniques are most effective for applying clustering algorithms aimed at identifying learning behavior patterns in ULE learning sessions?
- RQ2: Which clustering algorithm is most effective for identifying learning behaviour patterns and finding distinct groups in ULE learning sessions?

The results indicate that feature selection and aggregation are the most effective techniques for constructing the dataset, improving pattern detection in student click-based learning sessions. Additionally, while K-Means achieved the best overall performance, Agglomerative Clustering and HDBSCAN showed advantages under specific data transformations that incorporate contextual attributes.

The identified groups revealed statistically significant differences in most of the attributes analyzed and showed evidence of self-regulated learning behaviors, particularly among students who most frequently accessed the platform's main educational resources, such as lesson tools, collaboration spaces, and quizzes, and demonstrated indicators of success in these activities.

This paper is structured as follows: Section 2 presents the background concepts and related works; Section 3 describes the materials and methods used in this study; Section 4 discusses the results obtained; and, finally, Section 5 presents the final considerations and possible directions for future research.

2 Background

This section presents the theoretical foundation of this study, focusing on Virtual Learning Environments (VLE), Educational Data Mining (EDM), Self-Regulated Learning (SRL), and related work in these areas. The concepts discussed here are fundamental to understanding the state of the art and provide the necessary context for this study.

2.1 Virtual Learning Environments

Virtual Learning Environments (VLEs) are digital platforms designed to support and enhance the teaching and learning process across different educational modalities, including face-to-face, hybrid, and distance education (Moore, 2013). Currently, such environments are widely employed in various educational contexts, covering all levels of education, from basic education to higher education and postgraduate studies. VLEs enable the integration of multiple content formats, such as texts, videos, discussion forums, and interactive activities, providing a more flexible, accessible, and personalized learning experience. In this way, students can access educational resources and interact with teachers and peers anytime and anywhere, which significantly contributes to the democratization of access to quality education.

To meet contemporary pedagogical demands, VLEs offer a wide range of functionalities, such as discussion forums, virtual classrooms, quizzes, practical activities, digital libraries, and assessment tools. However, the advancement of educational technologies has enabled the development of even more sophisticated environments, known as Smart Learning Environments (SLEs). These intelligent environments are designed to provide students with personalized resources, including guidance, support tools, recommendations for learning materials, and study strategy suggestions, in a manner that is contextualized to consider the place, time, and specific needs of each learner. Thus, SLEs go beyond functioning as mere repositories of digital content, establishing themselves as dynamic systems that enhance the learning process through the application of emerging technologies and the analysis of educational data (Kinshuk et al., 2016).

Within this context, Adaptive and Intelligent Educational Systems (AIESs) stand out as a significant advancement over traditional VLEs. These systems are designed to continuously monitor students' academic performance and autonomously adapt to their individual needs, adjusting the level of difficulty of activities, recommending new content, and proposing personalized pedagogical strategies (Brusilovsky & Millán, 2007). In this way, AIESs enable a more effective, student-centered learning experience, fostering the development of skills and competencies in an individualized, dynamic manner that aligns with the contemporary demands of education.

In this scenario, educational environments, when systematically utilized, are capable of generating a massive volume of data related to students' interactions, academic performance, and the use of available learning resources. These data constitute a valuable source of information that can be explored through analytical techniques to understand how students learn, identify behavioral patterns, and evaluate the contribution of educational resources to improving the teaching and learning process (R. S. J. D. Baker et al., 2016).

2.2 Educational Data Mining

Educational systems are great providers of data that can be used to discover relevant knowledge through the use of Data Mining (DM) techniques, which are part of the Knowledge Discovery process in Databases (KDD) and consist of extracting knowledge from large databases; in the context of learning data, this process is usually called Educational Data Mining (EDM). Considering the growth in the use of VLEs, both in Distance Learning (EaD) and as support for face-to-face or hybrid education, large amounts of data have been generated, including user interactions within the

system, discussion forums, evaluations, assessments, and other activities, which further expands the possibilities for applying EDM techniques to enhance educational processes.

In this scenario, the research field of EDM appears to intend to apply and/or develop methods to analyze data generated from educational environments. According to R. S. J. D. Baker et al. (2016), exploring such data is important to better understand students, how they learn, and the context in which learning occurs, in addition to other factors that can influence learning. In the context of Computers in Education, the use of data-driven approaches is strongly supported by the Evidence-Based paradigm, which advocates that educational decisions and interventions should be grounded in empirical data and scientific evidence. In this sense, data analysis techniques, such as those employed in EDM, play a fundamental role in generating reliable evidence about learning processes, reinforcing the importance of EDM as a key approach for understanding and improving educational outcomes (Bittencourt & Isotani, 2018). The discovery of knowledge from educational data involves several steps that can be summarized into three main phases, preprocessing, application of data mining techniques, and post-processing, as highlighted in Figure 1, which illustrates the overall EDM process and its stages.

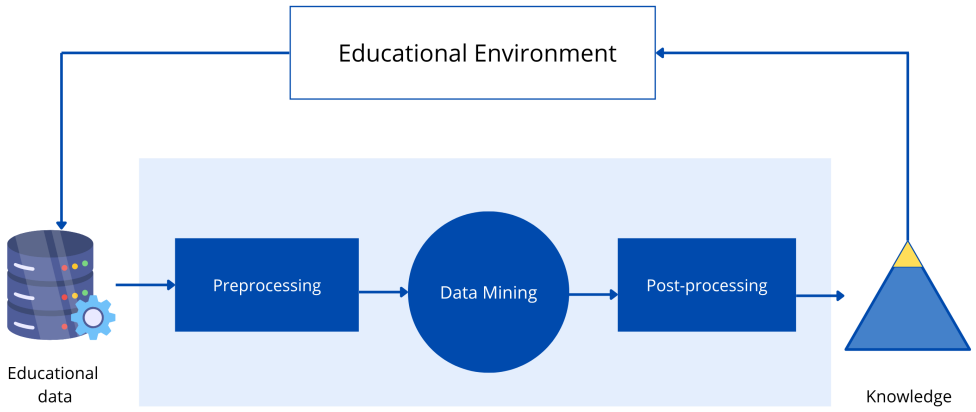


Figure 1: Educational Data Mining Process adapted from García et al. (2011).

One of the most important steps in the knowledge discovery process is data preprocessing, which consists of adjusting the obtained data to a more suitable format for the EDM process. This preparation aims at improving the mining process, and there are different techniques to do so, such as cleaning, selection, and transformation of attributes, aggregation, and manipulation of missing values, among others.

To select attributes from a dataset, for example, it is common to first analyze the normality of the data and the correlations among these attributes. For examining such correlations, statistical tests like Pearson or Spearman can be applied (Urdan, 2010). These tests help determine the degree of interdependence between two or more variables, revealing how the variation of one variable is related to changes in another.

The results of these analyses are expressed through correlation coefficients, which quantify the strength and direction of the relationship between variables. The Pearson coefficient measures the correlation between two quantitative variables that have a linear relationship, while the Spearman coefficient can be used even when the relationship is not linear and the variables are

not strictly quantitative. In both cases, the coefficients assume values within the interval [-1, 1], indicating whether the correlation is positive or negative.

Then, the central step of the knowledge discovery process is to apply mining techniques to the preprocessed data (García et al., 2011). In this step, the most appropriate technique is chosen for preprocessed data, such as classification, regression, association rules, groupings, sequential pattern mining, and text mining, among others. Finally, the final step in the EDM process, and no less important, is to interpret the results considering the context in which the data were collected to assist the decision-making process in the SLE. The interpretation and analysis of the results involve both specialists in the educational field and statistical tests to assess their significance.

A review conducted by Aldowah et al. (2019) highlights the main DM techniques used in the EDM process between the years 2000 and 2017. This study highlights that the most used techniques are classification and clustering, respectively. Classification is a supervised learning technique in which a predictive model is trained from a set of data that has input and output labels. In the educational setting, this type of learning can be used, for example, to predict student performance based on educational, personal, and social data (Devasia et al., 2016; Saa, 2016).

Clustering, or grouping, is a type of task considered as unsupervised learning, in which the dataset used has no labels; that is, the output of each record is not known. In this context, this type of learning can be applied in the educational scenario, for example, to find student learning routes (Bogarín Vega et al., 2016), analyze and examine the students' learning processes (Cerezo et al., 2016), or recommend groups to carry out collaborative activities (Monteverde et al., 2017).

Post-processing of data is a critical final step in the EDM process. After a technique is applied to the pre-processed data, the results must be interpreted to support decision-making within the educational environment. This interpretation and analysis require collaboration between educational specialists and the application of statistical tests to ensure the results are significant.

To assist the process of analyzing the results in the post-processing stage, it is important to carefully check which type of statistical significance test to use in each situation. In the case of clustering, for example, to compare the groups obtained by an algorithm, it is important to take the result of the normality test into consideration (Urdan, 2010). Statistical tests are essential to verify whether there are significant differences between the results obtained; furthermore, it is crucial to carry out a qualitative analysis of the results in order to gain a deeper understanding of the information related to students' learning processes and self-regulated behavior within these environments (R. Baker et al., 2020).

2.3 Self-Regulated Learning

Self-Regulated Learning (SRL) is a research area of Educational Psychology that studies personal aspects of students that influence their self-guided learning process. SRL is a conceptual framework for understanding the cognitive, meta-cognitive, behavioral, motivational, and emotional/affective aspects of learning (Panadero, 2017). In competitive and evaluative contexts, human achievements depend very much on the individual's ability to self-regulate (B. Zimmerman & Martinez-Pons, 1986).

Various SRL models are discussed in the literature. SRL models can be defined as cyclical and have different phases and sub-processes of self-regulation (Panadero, 2017; Pintrich, 2000;

Puustinen & Pulkkinen, 2001; B. J. Zimmerman, 1986). Although the models present different nomenclatures for the processes, their understanding allows them to be grouped into three major phases: a) Preparatory (or planning); b) Execution; and c) Evaluation.

The preparatory phase comprises the analysis of tasks, planning, definition of objectives, and establishment of goals (Panadero, 2017). In this phase, prior knowledge about a certain topic can be estimated, as well as activation of motivational beliefs, i.e., perceived self-efficacy, the value placed on the task, and personal interests. Planning the time and effort to be used for the task is also important at this stage. Administrative tools, calendars, and other technologies can be used to help students to plan the course development.

The second phase presented in the SRL models is called the execution phase, where tasks are performed while monitoring progress and performance (Panadero, 2017). According to B. J. Zimmerman and Moylan (2009), the two main processes during the performance are self-observation and self-control. Different strategies can be used to achieve these processes, for example, recording the time students spend reading a text would help them understand their individual profiles in the reading process, as well as highlighting excerpts from the text would help them to identify and recall the parts they consider most important. The work of Kitsantas (2013) describes several technologies that can help in this phase, such as social networks, virtual environments, administrative tools, testing tools, discussion forums, and bookmarks.

The evaluation phase is related to when students reflect, regulate, and adapt their learning process for future cycles (Panadero, 2017), which shows that learning is a dynamic and cyclical process. Self-reflection is built on self-judgment and self-reactions. At the end of a task, each student evaluates his/her learning process and points out possible causes for the achieved results. This kind of analysis can generate positive or negative emotional reactions. The use of multiple-choice questions or specific notes that appeared on the discussion forum are examples of tools that can be used at this stage.

In this context, educational systems generally have the potential to offer instructional and pedagogical architectures that take into account students' characteristics, such as levels of autonomy, learning styles, and learning pace, enabling them to act as active agents in their learning process. Research indicates that self-regulation is directly related to academic performance (B. J. Zimmerman, 1986). However, for this to occur effectively, it is essential that such systems not only provide resources to support SRL but also act proactively to foster the development of self-regulatory strategies (Melissa Ng Lee Yen, 2020). Furthermore, by mediating interactions and learning activities, these systems generate relevant data that can be analyzed to understand student behavior better and continuously improve pedagogical practices.

2.4 Related Works

Several studies have been conducted in the field of EDM to analyses student behavior in AVAs and understand the learning process. Dol and Jawandhiya (2023) presents a review of works published between 2010 and 2022, highlighting various EDM techniques applied in this context, such as classification, clustering, association rules, and time series. This review points out that the classification technique is the most used in the analyzed articles, followed by the combination with data clustering algorithms to assess student behavior, predict their academic performance, and investigate possible cases of school dropout.

With regard to the analysis of self-regulated behavior, specifically, there has been a significant increase in studies using EDM techniques in recent years. Recent research employs these techniques to investigate SRL from tracking data in AVAs. In particular, supervised and unsupervised learning techniques, such as clustering, have been applied to identify groups of students with different levels of SRL in online virtual environments (Damayanti et al., 2023).

The work described by Carmo et al. (2019) presents an approach based on sequential pattern mining to identify behaviors in the logs of an adaptive web system. In this investigation, the authors detected patterns of behavior in different categories, identified the most accessed resources, the most common types of navigation trajectories, and the frequency of use of materials. However, they point out that it was not possible to establish a clear relationship between the most used resources throughout the student's journey and their final performance.

In the same context of student behavior exploration, El-Halees (2009) applies four data mining techniques (association rules, classification, clustering, and outlier detection) to a dataset to examine student interaction and its influence on academic performance. The authors identified characteristics that impact results, behavior patterns prone to failure, clusters of students with similar attitudes, and anomalies in participant behavior.

Similarly, Lallé and Conati (2020) uses association and clustering rules to analyze student behavior in video consumption in massive open online courses (MOOCs). By clustering interaction data, the researchers found behavior profiles that favour or hinder effective learning and then extracted specific behavior patterns for each group using association rules.

In the field of self-regulated learning (SRL), the investigation by Ramos et al. (2020) analyses Moodle records to identify interaction profiles in a distance learning course. Hierarchical and non-hierarchical clustering methods were applied, resulting in the segmentation of students into three distinct profiles: low, medium, and high interaction with the virtual environment.

Moving on to more recent studies focused on self-regulation strategies, Rodriguez et al. (2021) explores self-regulated learning through the analysis of video clicks and study time management. Using the k-means algorithm, four patterns of self-regulation were identified, revealing that planning of activities is related to better academic performance.

Along the same lines, Farida and Sudibyo (2022) investigates the connection between self-regulation and academic performance by forming clusters using the k-means algorithm. Three distinct groups were identified (low, medium, and high levels of self-regulation), showing a positive correlation between greater self-regulation and better grades.

Finally, Peraić and Grubišić (2023) analyses engagement patterns in an Introduction to Programming course over three years, using Moodle records. Using k-means clustering, two main groups were identified: students with high engagement and good performance, and those with low engagement and poor performance.

While previous studies mainly focus on analyzing tracking data from the student's perspective, our research extends this scope by examining records from a ubiquitous educational environment, emphasizing learning sessions. Another important difference from related works is the systematic exploration of diverse data transformation and normalization techniques. By empirically testing how clustering algorithms respond to these variations, we emphasize the crucial role of preprocessing in uncovering meaningful behavioral patterns and identifying the most suitable

strategies for student behavior analysis. To provide a clearer comparison of the related studies and highlight their main characteristics, Table 1 summarizes the techniques employed, data sources, research focus, and key aspects of each work, while also emphasizing the differences between previous studies and the present research.

3 Materials and Methods

The material analyzed in this study consists of data on how students interacted with a ULE called Classroom eXperience (CX) (Cattelan et al., 2025). This particular ULE was designed around the Capture & Access (C&A) model, which follows the five stages initially proposed by Abowd et al. (1997) and later extended by Pimentel et al. (2001): *(i)* pre-production, focused on preparing content; *(ii)* live recording, which captures activities in real time; *(iii)* post-production, responsible for synchronizing and storing media streams in various formats; *(iv)* access, which makes the material available to users; and *(v)* extension, in which previously captured material can be enriched by users, as implemented in CX platform. Figure 2 illustrates this process.

Within this environment, there is a learning platform that enables participants to make annotations on lectures using an electronic whiteboard and revisit this material afterward. Furthermore, it offers tools for recommending and personalizing structured learning resources, in addition to social and collaborative features that contribute to enriching both the content and the overall learning experience.

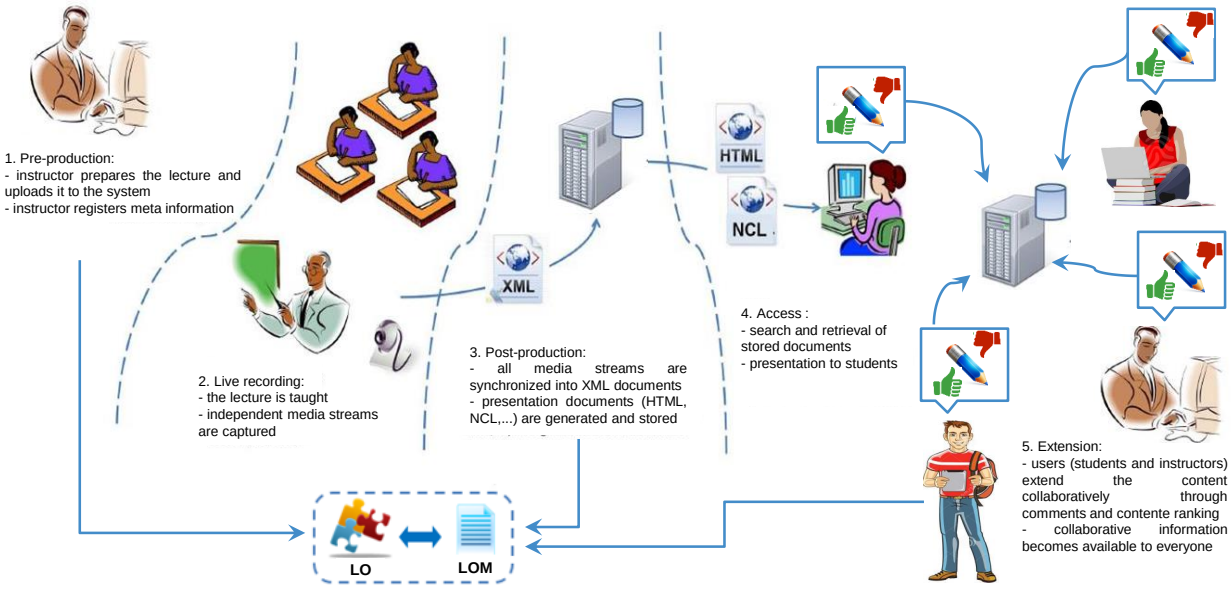


Figure 2: Five phases of the C&A process (Araújo et al., 2016).

The system's architecture includes a module dedicated to processing activities, which records interactions occurring within the environment and stores them in a repository Cattelan et al., 2025. Within this platform, a total of 23 distinct types of interactions are tracked as students access con-

Table 1: Comparative analysis of related works.

Study	Techniques	Data Source	Focus	Key Characteristics
Dol and Jawandhiya (2023)	Classification, clustering, association rules, time series	Multiple studies (review)	EDM techniques overview	Highlights classification as most used; combination with clustering
Damayanti et al. (2023)	Supervised and unsupervised learning (clustering)	AVA tracking data	SRL analysis	Identification of student groups with different SRL levels
Carmo et al. (2019)	Sequential pattern mining	Adaptive web system logs	Behavior patterns	Navigation trajectories and resource usage
El-Halees (2009)	Association rules, classification, clustering, outlier detection	Dataset of student interactions	Performance and behavior	Identification of patterns, clusters, and anomalies
Lallé and Conati (2020)	Clustering, association rules	MOOC video interaction data	Behavior in video consumption	Behavior profiles and patterns
Ramos et al. (2020)	Hierarchical and non-hierarchical clustering	Moodle logs	Interaction profiles	Three student profiles (low, medium, high)
Rodriguez et al. (2021)	K-means clustering	Video clicks and time management data	SRL strategies	Four SRL patterns related to performance
Farida and Sudibyo (2022)	K-means clustering	Student data	SRL and performance	Three SRL levels with positive correlation to grades
Peraić and Grubišić (2023)	K-means clustering	Moodle records	Engagement patterns	Two groups: high vs. low engagement
This study	Clustering algorithms with multiple preprocessing techniques	Ubiquitous learning environment records	Learning sessions	**Focus on learning sessions; systematic use of diverse data transformation and normalization techniques**

tent provided by instructors. These interactions can be categorized according to characteristics that represent different features of the environment, such as accessing classes, engaging in social and collaborative activities, performing self-assessments, searching for additional content, and personalizing content, as detailed in Table 2. Furthermore, each interaction record includes seven contextual attributes: four automatically identified (date and time of access, device type, screen resolution, and bandwidth) and three manually provided (reason for access, available time, and access location).

Table 2: Types of interaction captured by the CX platform (Costa et al., 2020).

Categories	**Interactions**
Access to class	Log in the environment; Lecture opening and closing; Slide navigation
Social and collaborative activities	Creation and exclusion of content bookmarks; like and dislike of bookmarks; indication of learning resource type; slide classification; comments and responses to comments (both in the context of lecture slides and in the context of a course)
Self-assessment	Answers to quizzes; changing of students' open model visualization
Search for additional content	Access to recommended additional content
Content Personalization	Responses to the instrument for assessing learning styles; changing the personalized content visualization

To better illustrate these interactions, Figure 3 shows a screenshot of a lecture captured by the teacher and already made available to students through a web interface. Item (1) indicates the slide navigation component. Each time the student changes slides, this type of interaction is stored. Item (2) highlights the bookmarking component, in which students can indicate the beginning of new content/subject (a subset of slides) inside a set of slides. Once created, the bookmark will be distributed to other students enrolled in the class, who may like, dislike, or create a new one. Only the owner of a bookmark can delete it. Item (3) shows a drop-down list to collaboratively indicate the types of educational resources in a slide based on the IEEE Learning Object Metadata specification (IEEE, 2002). Item (4) indicates the place where the student can rate the slide by giving one to five stars. The general average is shown in the star located in the upper right corner. Item (5) shows the comments component, in which students and teachers can interact with each other by inserting additional explanations, providing links to references and new material, asking questions, and pointing out important issues while accessing the content. Comments are eligible for replies and ratings, which encourage debate and can be used to measure their relevance. Item (6) shows the quiz component, where the student can answer multiple-choice questions created by the teacher specifically for the content studied in that set of slides.

To demonstrate how these interactions take place, Figure 3 presents a screenshot of a lecture recorded by the instructor and already accessible to students via a web interface. Item (1) marks the slide navigation tool, which records an interaction every time a student switches slides. Item (2) highlights the bookmarking feature, which allows students to mark the start of new content or

topics (a subset within the slides). Once a bookmark is created, it becomes visible to other students enrolled in the course, who can like, dislike, or create their own bookmarks. Only the student who created a bookmark can remove it. Item (3) points to a drop-down menu used collaboratively to classify the types of educational resources present on a slide, following the IEEE Learning Object Metadata standard (IEEE, 2002). Item (4) shows the area where students can rate a slide using a scale from one to five stars, with the overall average displayed in the star icon at the top right corner. Item (5) illustrates the comments section, where students and instructors can exchange ideas by adding clarifications, sharing links to external references or new materials, posting questions, and highlighting key points while navigating the content. Comments can receive replies and ratings, fostering discussion and helping gauge their usefulness. Lastly, item (6) displays the quiz area, where students can answer multiple-choice questions prepared by the instructor specifically for the material covered in that slide deck.

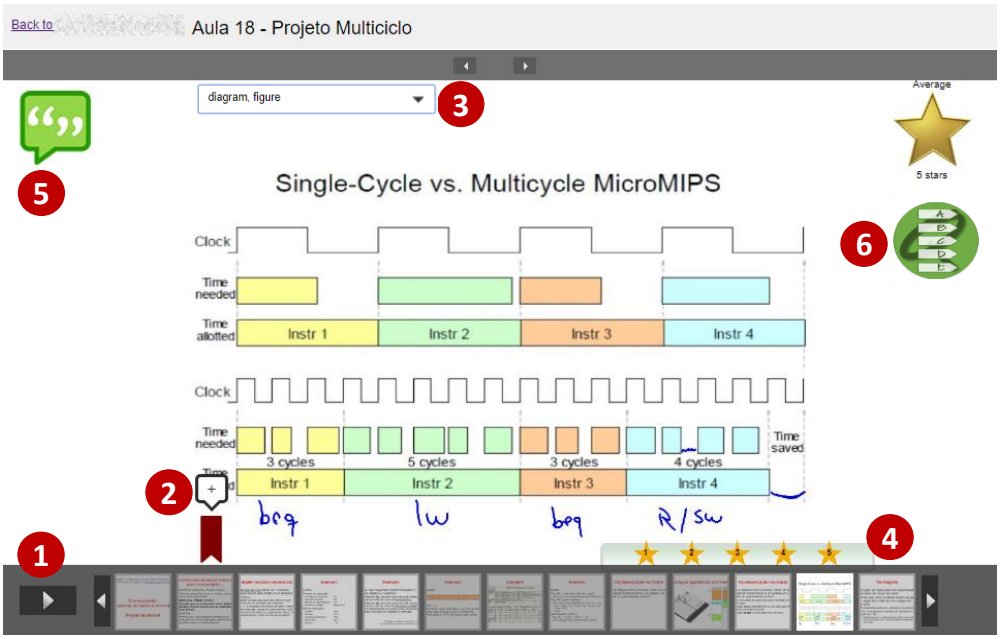


Figure 3: Screenshot of a lecture captured and made available on the CX platform (Costa et al., 2020).

To conduct the investigation and analysis of student behavior within the CX platform, the methodological process of this study was structured into five stages, as illustrated in Figure 4. In the first stage, student interaction records were extracted and segmented into distinct learning sessions, based on system access logs. Subsequently, an exploratory data analysis was performed to identify and select the most relevant attributes, as well as to construct multiple datasets for the next stage, which involved the application of clustering algorithms. Finally, the resulting clusters were analyzed to identify behavioral patterns in student profiles during learning sessions and to assess their implications for the teaching and learning process.

3.1 Data Gathering

The data for this study were gathered over the years 2017 and 2018 from 22 courses offered at the Federal University of Uberlândia, a Brazilian public higher education institution. This study was conducted in accordance with ethical standards for research involving human subjects and

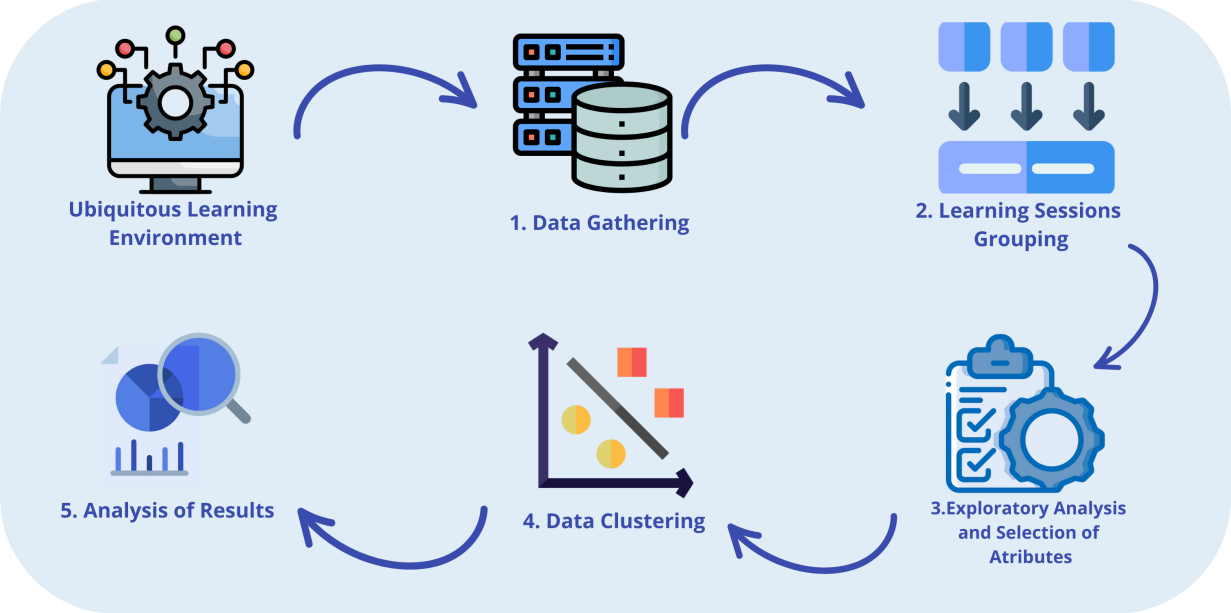


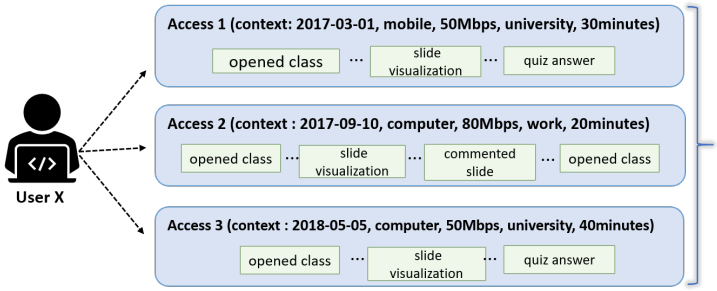
Figure 4: Overview of the working method.

received prior approval from the Ethics Committee for Research with Human Subjects under protocol number 46909515.4.0000.5152. It is important to note that the dataset reflects the profile and context of students enrolled in a specific public university, which may introduce limitations regarding the generalizability of the findings to other educational contexts, such as private institutions or different learning environments.

To extract the information, SQL (Structured Query Language) queries were performed, resulting in the retrieval of 3,187 access logs and 193,034 interaction records from the platform. Each recorded interaction is classified according to the user's action type, as detailed in Table 2. Furthermore, while the dataset provides detailed interaction records within a ubiquitous learning environment, caution should be taken when extending the results to other types of environments, such as traditional distance learning or hybrid educational settings.

3.2 Learning Sessions Grouping

Considering the tables generated during the data collection stage, the sessions were grouped using SQL queries. This step consisted of organizing each session according to the corresponding number and type of interactions performed. The interactions were aggregated into learning sessions, generating one or more records per user, each containing session-specific attributes. Figure 5 illustrates the process of grouping learning sessions by user. A learning session is defined as a user's access period within the system, determined by the time interval between login and logout. During each session, the user could perform various types of interactions, with unrestricted navigation; for example, it was not mandatory to view all available slides. Within this context, the grouping process aimed to determine the frequency of each interaction type performed during a session.



access	open class	slide visualization	slide coments	quiz answer
Access 1	1	1	0	1
Access 2	2	1	1	0
Access 3	1	1	0	1

Figure 5: Examples of interactions grouped into learning sessions adapted from Costa et al. (2020).

At the end of this stage, a file in CSV (Comma Separated Values) format was generated, containing 26 attributes representing 3,187 learning sessions. This file includes seven attributes that describe the access context (*reason*, *location*, *device*, *bandwidth*, *availabletime*, *accessdate*, and *accesstime*) and nineteen attributes that quantify the different types of interactions performed during the observed period.

Additionally, the attributes representing the category *Social and Collaborative Activities*" (see Table 2) were combined into a single attribute named "*collaboration*", since analyzing them separately resulted in numerous records with null values. Consequently, the dataset, initially comprising 26 attributes, was reduced to 15 attributes as detailed in Table 3.

3.3 Exploratory Analysis and Selection of Attributes

In this step, we initially analyzed the *accesstime* attribute. We found that some learning sessions (N = 5) had very long access times (between 6 and 11 hours) and few interactions. These cases occurred when students accessed the system and left the session open indefinitely. Therefore, we decided to update the time of these sessions to the average value of the other sessions.

Subsequently, from the original dataset containing 15 attributes and 3187 records, five alternative datasets were generated. It is important to note that these datasets were created independently from the original dataset, rather than sequentially.

The first dataset (D1) was obtained by removing attributes (columns) that contained missing values, resulting in a dataset without missing data while preserving all 3187 records.The second dataset (D2) was also generated by removing attributes with missing values. Additionally, the data were aggregated by *iduser*, resulting in 315 records (one per student). In this aggregation, the mean was used for numerical attributes and the mode for categorical attributes.

The third dataset (D3) was created without removing any attributes. Instead, missing values were handled through imputation, using the mean for numerical attributes and the mode for categorical attributes, maintaining the original 3187 records. The fourth dataset (D4) followed a similar approach to D3, where missing values were imputed using the mean and mode. However, in this case, the data were subsequently aggregated by *iduser*, resulting in 315 records. Finally, the fifth dataset (D5) was generated by removing specific attributes with a high proportion of missing values or low interaction relevance. This selection was supported by expert analysis, leading to

Table 3: Description of attributes.

Attribute	**Type**	**Description**
iduser	numeric	student identifier
reason	categorical	optional access context attribute, which corresponds to the reason why the student is accessing the system; this attribute can receive the values *missed_lesson*, *ordinary_study*, *quick_review* or *test*
location	categorical	optional access context attribute indicating the student's access location; this location can contain the values *home*, *university* or *work*
device	categorical	access context attribute automatically captured by the system that corresponds to the device used (*desktop* or *smartphone*)
bandwidth	numeric	access context attribute automatically captured by the system that corresponds to the bandwidth used to access the system
availabletime	numeric	optional access context attribute that corresponds to the access time in seconds that the student believes he/she will use in the session
accessdate	categorical	context attribute automatically captured by the system that corresponds to the date of access
accesstime	numeric	context attribute that corresponds to the actual access time in seconds (from *login* to *logout*)
lectureopen	numeric	the number of times the student opened a lesson during the session
slidevisualization	numeric	number of times the student viewed a slide
ilsanswer	numeric	the number of times the student responded, in that session, to the learning styles measuring instrument
changechart	numeric	number of times the student's open model view was changed in that session
collaboration	numeric	number of times the user has performed a collaboration on the system
quizanswer	numeric	corresponds to the number of *quizzes* that the student answered in that session
c_quizanswer	numeric	attribute that indicates how many *quizzes* the student got right

the exclusion of the attributes *reason*, *location*, *availabletime*, and *ilsanswer*, while preserving the 3187 records.

Next, an exploratory analysis was conducted on the five datasets. Initially, the Kolmogorov-Smirnov test was applied to identify whether the data followed a normal distribution or not and it was found that no attribute has a normal distribution in the five databases. Since the data do not follow a normal distribution, Spearman's coefficient was applied to the numerical attributes to analyze the correlation between these attributes. We identified a strong correlation between

the attributes *lectureopen* and *slidevisualization* (0.823), so we chose to exclude one of these attributes to eliminate possible redundancies. To decide which of the two attributes should be excluded, we analyzed the variance measure of the two attributes (*lectureopen*=6.2 and *slidevisualization*=8,411) and decided to exclude the *slidevisualization* attribute, since it had a higher variance value.

Then, it was necessary to transform the categorical attributes into numerical ones using the binarization technique. For datasets D1, D2, D3, and D4, the *reason* and *location* attributes were transformed using 1-of-n encoding, while the *device* attribute was transformed using binary integer encoding. In the case of dataset D5, only binary integer encoding was applied to the *device* attribute, since the other categorical attributes had already been excluded in the previous step. All pre-processing procedures were performed using SPSS software (IBM Corp. Released 2011, 2011). As a result, the datasets were generated and prepared for the application of the clustering algorithms, with their main characteristics summarized in Table 4.

Table 4: Number of records and attributes of each built dataset.

dataset	Number of records	Number of attributes
D1	1507	17
D2	315	17
D3	3187	17
D4	315	17
D5	3187	8

3.4 Data Clustering

There are various types of clustering algorithms, which can be categorized according to the approach used to form clusters. The main categories include partitioning, hierarchical, and density-based algorithms, each with its own specific characteristics and clustering strategies. In this study, we selected three algorithms, each representing one of these categories, to analyze the dataset.

The first algorithm chosen was Agglomerative Clustering, which belongs to the family of hierarchical algorithms and employs a bottom-up approach. In this method, each data point initially forms its own cluster, and at each iteration, the closest pairs of clusters are merged (Tan et al., 2016). The key feature of this approach is the generation of a cluster tree, known as a dendrogram, which is useful for uncovering the hierarchical structure within the data.

The second algorithm selected was HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a divisive hierarchical clustering algorithm based on the Minimum Spanning Tree (MST) derived from the Mutual Reachability Distance Graph. HDBSCAN extends the density-based DBSCAN algorithm to a hierarchical context (Campello et al., 2013), addressing some of DBSCAN's limitations, such as difficulties in handling clusters with varying densities. Additionally, it is effective at detecting and removing noise (outliers). The HDBSCAN does not require specifying the number of clusters in advance, as clusters are identified based on data density. However, it does require defining the minimum number of points to form a cluster. The algorithm works by progressively removing edges with the highest weights (highest mutual reachability distances) from the Minimum Spanning Tree, thereby establishing different hierarchy levels with connected groups and isolated noise points.

Finally, the third algorithm adopted was the classic K-Means (Hartigan & Wong, 1979), which is classified as a partitioning method. This algorithm divides a set X of n samples into k disjoint groups, each described by the mean u of the samples in the group, known as the centroid. The number of groups (k) must be defined in advance. The algorithm randomly initializes k sample points as centroids, and for each sample, it computes its distance to each centroid, assigning the sample to the group of the nearest centroid.

These algorithms were selected due to their diverse approaches and clustering strategies, enabling an assessment of which method performs best on the analyzed data and how each handles the various data transformations applied during the preprocessing stage of this study. The Agglomerative Clustering and K-Means algorithms were applied using the scikit-learn library[1], while the HDBSCAN algorithm was employed through its dedicated library[2].

3.5 Analysis of results

To assess the quality of the clusters generated by each algorithm, we first applied the internal validation metric silhouette coefficient, which measures how similar each point is to its cluster compared to other clusters. Higher values indicate better-defined and more cohesive clusters. After the analysis of the silhouette coefficient, we evaluated the best clustering results for each algorithm by examining the mean differences of each attribute within the identified clusters. To verify whether these differences were statistically significant, we used the Python SciPy[3] library.

Before applying any statistical test for variance analysis, it was necessary to verify the underlying assumptions regarding the distribution of the data. Therefore, the one-sample Kolmogorov–Smirnov test (Razali & Wah, 2011) was applied to assess whether the variables followed a normal distribution. The test was conducted at a significance level of $\alpha = 0.05$, and the null hypothesis of normality was rejected (p-value < 0.05). As a result, the non-parametric Kruskal–Wallis test (Urdan, 2010) was used to compare two or more independent samples and determine whether there were statistically significant differences among the attributes across the clusters.

4 Results

This section presents the results obtained by applying the clustering algorithms to the five datasets created with different preprocessing transformations. The aim is to identify which algorithms yield the best clustering performance and to analyze the impact of the various data transformation strategies on the clustering outcomes, thereby revealing which combinations of techniques produce the most cohesive and meaningful groupings for the behavior under study.

[1] https://scikit-learn.org/stable/

[2] https://hdbscan.readthedocs.io/en/latest/

[3] https://www.scipy.org/

4.1 Agglomerative Clustering

The Agglomerative Clustering Algorithm in the scikit learn library has 7 input parameters with default values, but in these tests, 3 parameters were modified: $linkage = single$, $affinity = euclidian$, and $n\_clusters$, which were given different values (2,3,4, and 5) to analyze the relative measure of the silhouette coefficient. Figure 6 shows the variation of this measure for five datasets in relation to the number of clusters. We can see from the graph that the silhouette measure is better for dataset D3, as it reaches values above 0.8 for 2, 3, and 4 clusters. However, when it comes to the agglomerative algorithm, assessing the quality of the clustering with the silhouette measure alone is not interesting, as it is necessary to assess whether the data has been clustered.

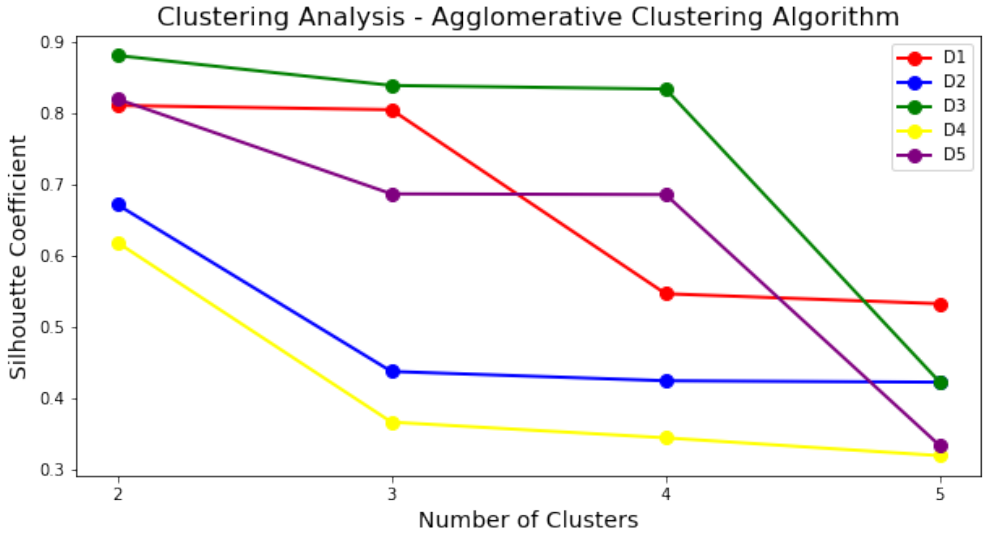


Figure 6: Variation of the Silhouette Coefficient with relation to the number of clusters for all five datasets - Agglomerative Clustering.

We analyzed the dendrograms generated by the five datasets, and it was observed that only for datasets D1 and D3, the agglomerative algorithm performed the clustering with cut points equal to 4 and 5 clusters, respectively. For the D1 dataset with $n\_clusters = 4$, we have a silhouette coefficient equal to 0.55, and we observed which two clusters ($C0$ with 887 records, $C1$ with 618 records) and two singletons. For the D3 dataset with $n\_clusters = 5$, we obtain a silhouette coefficient equal to 0.42 and observe two clusters ($C0$ with 2566 records, $C1$ with 618 records) and three singletons. Therefore, we consider that the Agglomerative algorithm performed best for the D1 dataset. Table 5 shows statistical details containing the mean, the standard deviation, and the ρ-value for each quantitative attribute of clusters emerged in the D1 dataset, which obtained the best silhouette coefficient value among all tests.

Additionally, we analyzed the effect size (Cohen's d) to assess the magnitude of the differences between clusters. The results indicate that most attributes present small effect sizes, suggesting limited practical differences between clusters despite statistical significance. However, the attribute *availabletime* exhibited an extremely large effect size ($d = -6.32$), indicating a substantial difference between clusters and highlighting its strong discriminative power. Moderate effects were also observed for *accesstime* ($d = -0.43$), while the remaining attributes showed small effects. These findings suggest that, although several variables are statistically significant, only a subset meaningfully contributes to distinguishing student behavior patterns.

Table 5: Statistics per cluster generated by Agglomerative in D1 dataset, the ρ-value for the Kruskal-Wallis test, and effect size (Cohen's d).

Attribute	***Cluster* C0 (N=887)** $\bar{x} \pm dp$	***Cluster* C1 (N=618)** $\bar{x} \pm dp$	**ρ-value**	**Cohen's d**
bandwidth	534.49 ± 460.65	641.16 ± 499.49	.00000	-0.22
availabletime	$1,724.92 \pm 1,136.10$	$7,238.93 \pm 28.67$	.00000	-6.32
accesstime	$1,346.61 \pm 2,542.81$	$2,569.60 \pm 3,222.68$	.00000	-0.43
lectureopen	1.71 ± 2.02	2.30 ± 2.99	.00437	-0.24
ilsanswer	0.23 ± 1.04	0.09 ± 0.40	.00005	0.17
changechart	0.08 ± 0.56	0.12 ± 0.74	.61967	-0.06
collaboration	0.69 ± 8.03	2.37 ± 14.95	.00440	-0.15
quizanswer	0.24 ± 2.66	0.55 ± 3.23	.00001	-0.11
c_quizanswer	0.15 ± 1.49	0.32 ± 1.78	.00001	-0.11

We observed that Cluster C1 presents higher means in all attributes, except the *ilsanswer* attribute. In addition, these means are statistically significantly higher according to the Kruskal-Wallis test, except the *changechart* attribute, with $p\_value = .61967$. However, considering the effect size analysis, most of these differences are of small magnitude, indicating limited practical significance for several attributes. Thus, we observe that the clusters found by the Agglomerative Clustering in the dataset D1 are statistically distinct, but only partially differentiated in practical terms, and cluster C1 contains longer learning sessions, with more lecture openings, more collaboration, and consequently more attempts to answer the quizzes (with more correct attempts too).

4.2 HDBSCAN

Density-based algorithms, such as HDBSCAN, find high-density regions that are surrounded by low-density regions. The discovery of groups is performed arbitrarily, not needing to inform the number of groups as parameters; however, other density-related parameters are required. According to Campello et al. (2013), the HDBSCAN has two input parameters:

- ***min_cluster_size (mcs):*** minimum number of elements required to form a cluster;
- **min_samples (ms):** parameter that directly controls the minimum size of groups, indicating how conservative a group should be.

The ***min_samples*** is the main parameter to estimate the density. Choosing the parameter value is not a trivial task. On the one hand, a very low value can result in the formation of spurious clusters or fragmentation into many small clusters, on the other hand, the higher the value of the min samples parameter, the more conservative will be the clustering so more points will be declared as *outliers* and the clusters will be restricted to progressively denser areas. The value definition for the min samples parameter was done empirically after several tests on the datasets. As the datasets present distinct characteristics, different values were set in each dataset. Figure 7 shows the values of the input parameters for each dataset and the silhouette coefficient. We can see that the best silhouette coefficient value was found in dataset D1.

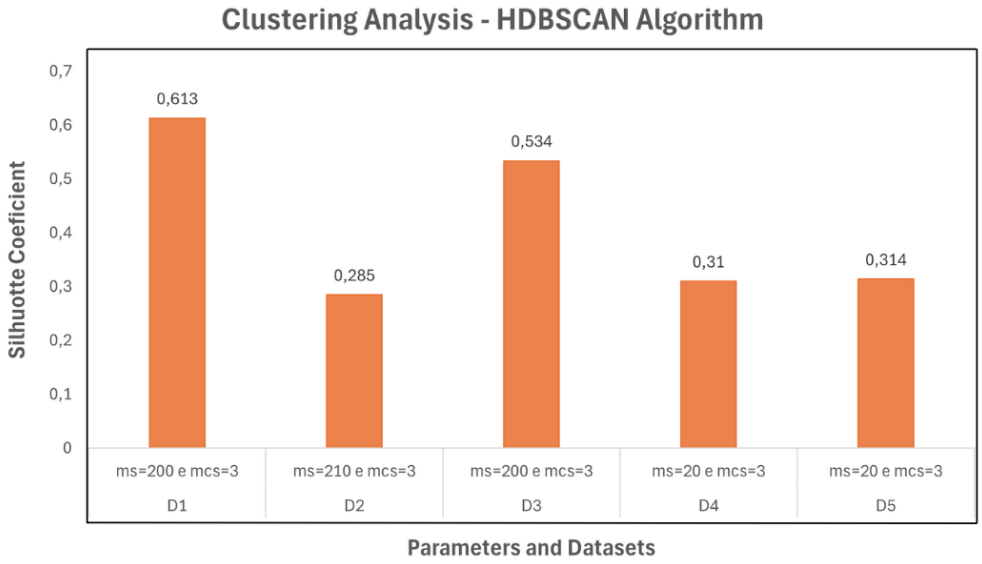


Figure 7: Input parameters of the HDBSCAN algorithm for each dataset and their respective silhouette coefficient value.

Table 6 presents a comparison of the mean values for the quantitative attributes in the dataset D1. In all attributes, *Cluster C*1 presented higher mean values, except *ilsanswer*. Regarding the Kruskal Wallis test, we can see that only two attributes did not present statistically significant differences: *lectureopen* with $p\_value = .11498$, and *changechart* with $p\_value = .93844$.

In addition, the effect size analysis (Cohen's d) indicates that most attributes present small effect sizes, suggesting limited practical differences between clusters despite statistical significance. However, the attribute *availabletime* shows an extremely large effect size ($d = -6.48$), highlighting a strong distinction between clusters. A moderate effect was also observed for *accesstime* ($d = -0.61$), while the remaining attributes exhibit small effects.

It can be observed that the *clusters* found in dataset D1 by the HDBSCAN algorithm are statistically distinct, but only partially differentiated in practical terms, with longer learning sessions in *Cluster C*1, with more lecture opening and attempts to *quizzes* - as well as with more correct attempts.

Table 6: Statistics per cluster generated by HDBSCAN in D1 dataset, the ρ-value for the Kruskal-Wallis test, and effect size (Cohen's d).

Attribute	***Cluster* C0 (N=801)** $\bar{x} \pm dp$	***Cluster* C1 (N=527)** $\bar{x} \pm dp$	**ρ-value**	**Cohen's d**
bandwidth	537.84 ± 449.43	641.02 ± 500.17	.00002	-0.22
availabletime	1665.16 ± 1107.78	7238.82 ± 28.70	.00000	-6.48
accesstime	653.45 ± 996.61	1457.98 ± 1696.36	.00000	-0.61
lectureopen	1.56 ± 1.80	1.95 ± 2.57	.11498	-0.18
ilsanswer	0.24 ± 1.09	0.09 ± 0.38	.00006	0.17
changechart	0.07 ± 0.54	0.09 ± 0.60	.93844	-0.04
collaboration	0.66 ± 8.19	1.97 ± 12.61	.04078	-0.12
quizanswer	0.22 ± 2.73	0.41 ± 3.01	.00008	-0.07
c_quizanswer	0.14 ± 1.50	0.22 ± 1.46	.00008	-0.05

4.3 K-Means

To obtain the ideal number of clusters, an analysis of the silhouette coefficient was performed with different values of k for each dataset. As can be observed in Figure 8, dataset D5 obtained better results for all values of k. After analyzing the clusters generated in each dataset, it can be observed that the value of $k = 2$ for the D5 dataset was the one that could best separate the data in terms of significant differences between the averages of each quantitative attribute and the silhouette coefficient ($S_i = 0.75$).

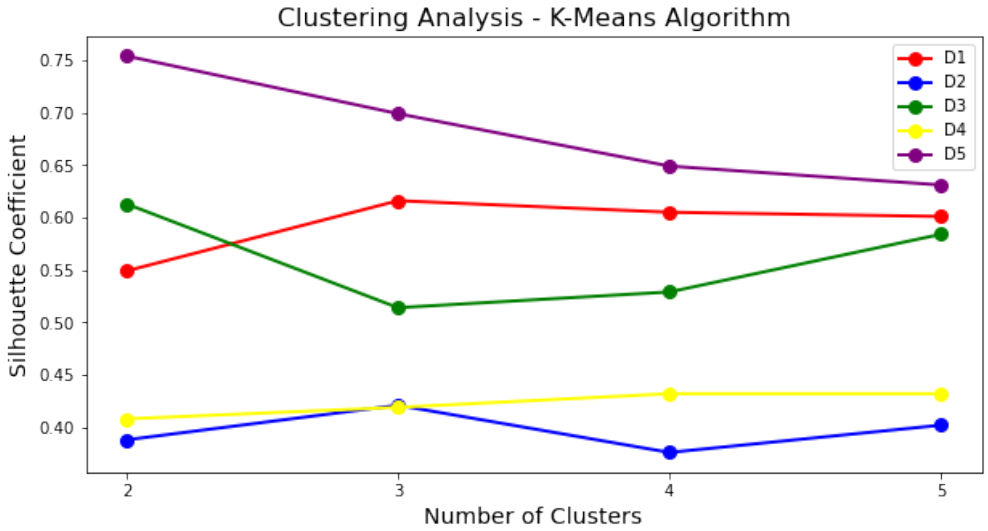


Figure 8: Variation of the Silhouette Coefficient with relation to the Number of Clusters for the five datasets - K-Means Algorithm.

Table 7 presents data from descriptive statistics containing the mean, standard deviation, and *p_value* of each attribute. To verify if there are statistically significant differences between the mean values of the attributes, Kruskal-Wallis's non-parametric test was performed, since the data does not come from a normal distribution, as confirmed by the Kolmogorov-Smirnov. Table 7 illustrates a comparison of these averages, and it can be seen that Cluster 1 has the highest average in all attributes, except bandwidth, with $\chi^2 = 612.47, \rho = .09696$. Therefore, statistical significance was confirmed by the Kruskall-Wallis test.

Table 7: Statistics per cluster generated by K-Means in D5 dataset, the ρ-value for the Kruskal-Wallis test, and effect size (Cohen's d).

Attribute	***Cluster* C0 (N=2774)** $\bar{x} \pm dp$	***Cluster* C1 (N=413)** $\bar{x} \pm dp$	**ρ-value**	**Cohen's d**
bandwidth	612.47 ± 485.14	574.72 ± 475.42	.09696	0.08
accesstime	780.32 ± 1142.64	8448.33 ± 4081.18	.00000	-2.56
lectureopen	1.65 ± 1.89	3.72 ± 3.74	.00000	-0.70
changechart	0.1 ± 0.67	0.30 ± 1.24	.00000	-0.20
collaboration	0.93 ± 15.54	1.38 ± 11.94	.00071	-0.03
quizanswer	0.33 ± 2.53	1.36 ± 4.25	.00000	-0.29
c_quizanswer	0.22 ± 1.50	0.96 ± 3.03	.00000	-0.31

The results presented in Table 7 reveal statistically significant differences between clusters for most variables. In addition to statistical significance, effect sizes were computed to assess the magnitude of these differences. The results indicate that accesstime presents a very large effect size, suggesting a substantial difference in temporal engagement patterns between groups. Furthermore, attributes such as lectureopen and slidevisualization show moderate to large effects, indicating meaningful differences in content interaction behaviors.

Based on these findings, it is possible to characterize distinct behavioral profiles. Cluster C1 presents higher values for attributes related to interaction and content engagement, such as access time, lecture access, and assessment-related activities, indicating a more active and consistent interaction pattern. This suggests that students in this group demonstrate behaviors associated with higher levels of self-regulated learning. In contrast, Cluster C0 shows lower levels of interaction across these attributes, reflecting a less engaged behavioral profile. Therefore, Cluster C1 can be interpreted as representing students with *high SRL behavior*, whereas Cluster C0 corresponds to students with *low SRL behavior*. These behavioral patterns provide actionable insights for the design of adaptive interventions in ubiquitous learning environments. For instance, students in the low SRL cluster may benefit from additional scaffolding and guidance, while those in the high SRL cluster may be supported through more advanced and autonomous learning activities.

Figure 9 highlights the differences between each attribute, showing the value of each attribute for each cluster. for example, the learning sessions in cluster 1 had higher average interactions across various resources. Sessions with longer access times and more collaborative activities also demonstrated a higher frequency of correct quiz answers, suggesting that these sessions may represent self-regulated behavior within this SLE. According to B. J. Zimmerman (2008), self-regulated learning involves how proactive students take responsibility for their learning process. That is, students who engage more in the process, through increased collaboration and interaction, tend to achieve better performance outcomes as they apply self-control techniques, set goals, and develop strategies to succeed in learning.

In this context, it is important to highlight that answering quizzes is not a mandatory assessment for the student; it is optional. Kitsantas (2013) emphasizes that "self-regulated students tend to self-assess frequently and objectively using self-monitored data", which again indicates evidence of self-regulated student behavior in this analysis. She also shows several surveys carried out to provide guidelines on how VLEs can be used to support students' self-regulation, specifically in higher education.

In addition, the author highlights the results of some empirical research aimed at adapting VLEs in this process, suggesting that when such environments are properly designed, they can improve students' self-regulation processes. She also suggests that students are more likely to be involved with self-regulated learning when educational technologies have tools that adequately support this process, which contributes to improving the learning process (B. J. Zimmerman, 1986). Therefore, it was clearly shown that the analyzed SLE in this work adequately supports the SRL process, as confirmed by the data analysis previously provided.

4.4 Discussion and Comparison of Results

Among the transformed datasets, D1 and D5 emerged as the most effective for identifying learning behavior patterns in ULE learning sessions. D1 produced consistent and relevant results across

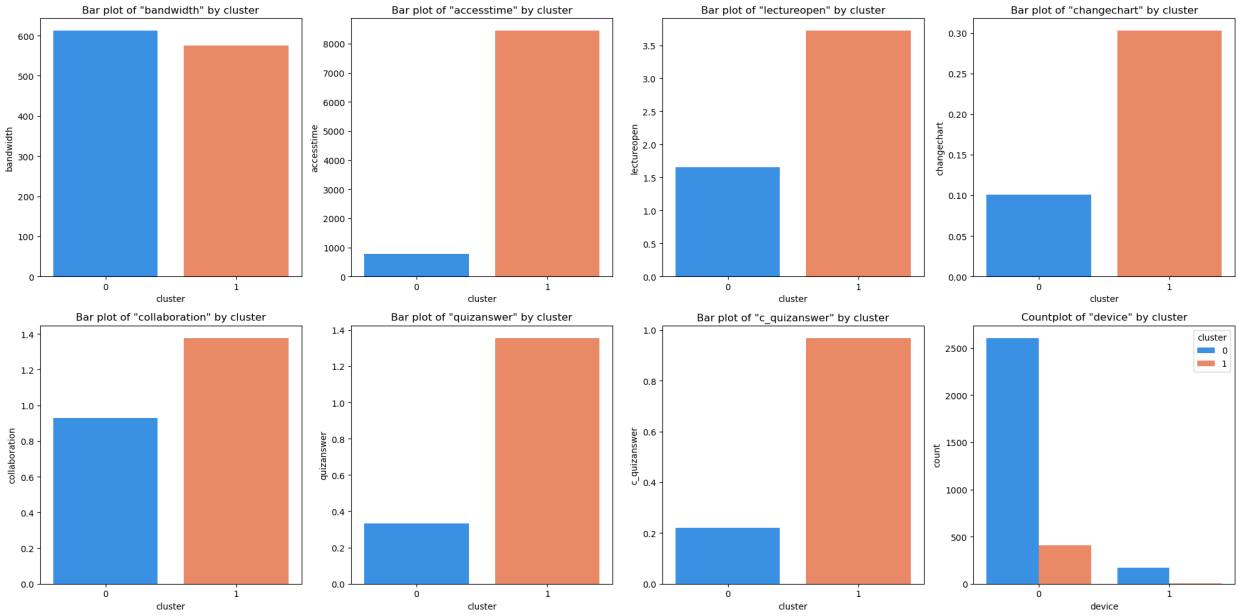


Figure 9: K-Means Algorithm - dataset D5 - Comparison of attributes in each cluster.

all three clustering algorithms tested. Although D3 also performed well with K-Means and HDBSCAN, its results with Agglomerative Clustering were less robust, as this algorithm did not produce statistically significant differences in the Kruskal-Wallis test for this dataset.

In the case of D1, the results obtained with HDBSCAN and Agglomerative Clustering were similar overall, differing only in the attribute *lectureopen* (number of times the student opened a lesson). For Agglomerative Clustering, this attribute showed a statistically significant difference between clusters, while for HDBSCAN it did not — likely due to HDBSCAN's tendency to classify some records as outliers (179 in this specific case).

Importantly, in response to RQ1, the D5 dataset, refined by a domain specialist to remove attributes representing resources that were rarely used, demonstrated superior performance compared to the other data transformations tested. This dataset achieved the best overall results, with the highest silhouette coefficient values and stronger statistical significance, although the practical differences were concentrated in a subset of key attributes. Notably, addressing RQ2, the K-Means algorithm, when applied to the D5 dataset, was the most effective among the clustering methods tested, consistently generating more distinct and interpretable clusters than HDBSCAN and Agglomerative Clustering.

All the algorithms produced clusters showing broadly similar patterns of learning behavior. However, considering the effect size analysis (Cohen's d), it is important to note that, although several attributes showed statistically significant differences across clusters, most of them presented small effect sizes, indicating limited practical significance. Across all algorithms, only a few attributes, particularly *availabletime*, and to a lesser extent *accesstime*, consistently exhibited moderate to large effect sizes, suggesting that these variables play a more substantial role in distinguishing learning behavior patterns.

Typically, one cluster indicated a more participatory learning pattern, characterized by longer study time and higher bandwidth availability, while the other reflected a less participatory pattern, albeit with a higher response rate to the learning styles index. However, this distinction is primarily driven by attributes with larger effect sizes, whereas other attributes contribute less substantially to differentiating these patterns.

These findings reinforce the crucial role of careful data pre-processing and demonstrate how different transformations can strongly influence the quality and interpretability of the clusters. In particular, the consistency of the results across different clustering algorithms strengthens the robustness of the identified patterns. The superior performance of D5, combined with the effectiveness of the K-Means algorithm, highlights that combining expert-driven data refinement with an appropriate clustering method is key to extracting meaningful patterns from the complex and diverse interaction data collected in ULE learning sessions. Future work should further investigate how the key attributes identified (e.g., *availabletime* and *accesstime*) relate to students' engagement, performance, and self-regulated learning strategies.

Beyond the immediate findings, these results open up several opportunities for extending the application of clustering-based analyses in ubiquitous learning environments. For instance, the identified behavioral patterns can support the development of adaptive dashboards that provide real-time feedback to both students and instructors, enabling more informed pedagogical decisions. In addition, the integration of these approaches with Open Educational Resources (OERs) may allow for the recommendation of personalized learning materials aligned with students' interaction patterns. Furthermore, future investigations should explore the applicability of these methods in diverse educational contexts, including elementary and middle school settings, in order to assess the generalizability of the findings across different age groups and learning scenarios.

5 Conclusion and Future Work

In this work, it was possible to observe the importance of educational data mining techniques for the analysis of students' behavior in virtual learning environments. Three data mining algorithms from different categories (partitional, hierarchical, and density-based) were tested on five different datasets produced with interactions collected from a Ubiquitous Learning Environment. Results showed that the algorithms are very sensitive to data transformation and that K-Means achieved a better result for the silhouette coefficient measure (75%) in one of the databases.

Additionally, although statistically significant differences were observed between clusters, the effect size analysis indicated that only a subset of attributes—particularly those related to study time and interaction—presented substantial practical differences. Also, two groups with statistically significant differences were found, confirmed by statistical tests. In this context, the clusters highlighted differences between attributes, showing that sessions with longer access times and more collaborative activities also resulted in more correct quiz answers.

These findings suggest that students with higher levels of collaboration and interaction tend to achieve better performance results, indicating behaviors potentially associated with self-regulated learning, such as goal setting and strategic planning. Thus, results provide evidence of

behaviors associated with SRL, which is a desirable characteristic in this type of learning environment. However, it is important to note that these conclusions are based on observed behavioral patterns and do not imply causal relationships.

In this context, it was possible to observe that students demonstrating behaviors associated with SRL performed better throughout learning sessions. Consequently, this work helps to reinforce the thesis that providing learning environments that foster self-regulated learning behavior is very important. In this process, students are protagonists of their learning and can develop several cognitive strategies, metacognition, motivation, and emotion/affection to self-regulate their learning. As previous research has also shown, students with self-regulated behavior tend to perform better during the learning process.

Therefore, this study contributes to the existing literature by demonstrating how different data preprocessing strategies and session-level analyses can influence the identification of meaningful behavioral patterns in ubiquitous learning environments, as well as by highlighting the importance of combining statistical significance with effect size analysis.

As future work, it is intended to apply other clustering algorithms, as well as other data mining techniques and statistical tests to analyze student behavior, both in learning sessions and at the course level. In addition, we want to conduct a more in-depth study of behavioral characteristics in the learning sessions that encourage or hinder self-regulated learning in this type of environment. Therefore, this analysis would adequately guide the development of new features in SLE, considering better support for high-quality learning processes.

Furthermore, future work should include the use of pre- and post-tests to better investigate potential causal relationships between self-regulated learning behaviors and academic performance. Future research may also explore the development of adaptive dashboards based on identified behavioral patterns, the integration with Open Educational Resources (OERs), and the application of these approaches in diverse educational contexts, such as elementary and middle school environments, in order to assess the generalizability of the findings.

Acknowledgements

The authors thank the Federal University of Uberlândia (UFU) and the Federal Institute of Education, Science, and Technology of South of Minas Gerais (IFSULDEMINAS) for their support in the development of this research.

References

Abowd, G. D., Atkeson, C. G., Feinstein, A., Hmelo, C., Kooper, R., Long, S., Sawhney, N., & Tani, M. (1997). Teaching and learning as multimedia authoring: The classroom 2000 project. *Proceedings of the Fourth ACM International Conference on Multimedia*, 187–198. https://doi.org/10.1145/244130.244191 [GS Search].

Aldowah, H., Al-Samarraie, H., & 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. https://doi.org/10.1016/j.tele.2019.01.007 [GS Search].

Araújo, R. D., Brant-Ribeiro, T., Ferreira, H., Dorça, F., & Cattelan, R. (2016). Segmentação colaborativa de objetos de aprendizagem utilizando bookmarks em ambientes educacionais ubíquos. *Simpósio Brasileiro de Informática na Educação*, 1205–1214. [GS Search].

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), 1–24. https://doi.org/10.1186/s41239-020-00187-1 [GS Search].

Baker, R. S. J. D., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics (A. A. Rupp & J. P. Leighton, Eds.), 379–396. https://doi.org/10.1007/978-1-4614-3305-7_4 [GS Search].

Bittencourt, I. I., & Isotani, S. (2018). Informática na educação baseada em evidências: Um manifesto. *Revista Brasileira de Informática na Educação*, *26*(03), 108. [GS Search].

Bogarín Vega, A., Romero Morales, C., & Cerezo Menéndez, R. (2016). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. *Edmetic*. https://doi.org/10.21071/edmetic.v5i1.4017 [GS Search].

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web: Methods and strategies of web personalization* (pp. 3–53). Springer. https://doi.org/10.1007/978-3-540-72079-9_1 [GS Search].

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14 [GS Search].

Carmo, Ê. P., Gasparini, I., & Oliveira, E. (2019). Captura e visualização das trajetórias de aprendizagem como ferramentas para a análise do comportamento dos estudantes em um ambiente adaptativo educacional. *Simpósio Brasileiro de Informática na Educação*, 309–318. https://doi.org/10.5753/cbie.sbie.2019.309 [GS Search].

Cattelan, R. G., Araújo, R. D., Ferreira, H. N., Brant-Ribeiro, T., & Dorça, F. A. (2025). Classroom experience: From automated multimedia capture to personalized learning. *Multimedia Tools and Applications*, *84*(24), 27609–27645. https://doi.org/10.1007/s11042-024-20238-3 [GS Search].

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 [GS Search].

Costa, J. A., Dorça, F. A., & Araújo, R. D. (2020). Avaliação do comportamento de estudantes em um ambiente educacional ubíquo. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 182–191. https://doi.org/10.5753/cbie.sbie.2020.182 [GS Search].

Damayanti, A., Kusumawardani, S. S., & Wibirama, S. (2023). A review of learners' self-regulated learning behavior analysis using log-data traces. *2023 IEEE 12th International Conference on Engineering Education (ICEED)*, 90–95. https://doi.org/10.1109/ICEED59801.2023.10264050 [GS Search].

Devasia, T., Vinushree, T., & Hegde, V. (2016). Prediction of students performance using educational data mining. *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167 [GS Search].

Dol, S. M., & Jawandhiya, P. M. (2023). Classification technique and its combination with clustering and association rule mining in educational data mining—a survey. *Engineering Applications of Artificial Intelligence*, *122*, 106071. https://doi.org/10.1080/03055698.2018.1516628 [GS Search].

El-Halees, A. M. (2009). Mining students data to analyze e-learning behavior: A case study. *Mining students data to analyze e-Learning behavior: A Case Study*, *29*. https://doi.org/10.1109/icca-ticet.2018.8726203 [GS Search].

Farida, A., & 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. https://www.academia.edu/download/90973232/5141.pdf [GS Search].

García, E., Romero, C., Ventura, S., & Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, *14*(2), 77–88. https://doi.org/10.1016/j.iheduc.2010.07.006 [GS Search].

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. Retrieved June 22, 2024, from http://www.jstor.org/stable/2346830 [GS Search].

IBM Corp. Released 2011. (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp. https://doi.org/10.1016/b978-0-12-815764-0.00027-7 [GS Search].

IEEE. (2002, July). Draft standard for learning object metadata. http://ltsc.ieee.org/wg12/files/LOM%5C_1484%5C_12%5C_1%5C_v1%5C_Final%5C_Draft.pdf [GS Search].

Kinshuk, Chen, N.-S., Cheng, I.-L., & Chew, S. W. (2016). Evolution is not enough: Revolutionizing current learning environments to smart learning environments. *International Journal of Artificial Intelligence in Education*, *26*(2), 561–581. https://doi.org/10.1007/s40593-016-0108-x [GS Search].

Kitsantas, A. (2013). Fostering college students' self-regulated learning with learning technologies. *Hellenic Journal of Psychology*, *10*(3), 235–252. https://pseve.gr/wp-content/uploads/2018/03/Volume10_Issue3_Kitsantas.pdf [GS Search].

Lallé, S., & Conati, C. (2020). A data-driven student model to provide adaptive support during video watching across MOOCs. *International Conference on Artificial Intelligence in Education*, 282–295. https://doi.org/10.1007/978-3-030-52237-7_23 [GS Search].

Melissa Ng Lee Yen, A. (2020). The influence of self-regulation processes on metacognition in a virtual learning environment. *Educational Studies*, *46*(1), 1–17. [GS Search].

Monteverde, I., Amaral, G., Ramos, D., Nascimento, P., Gadelha, B., & Oliveira, E. (2017). M-cluster: Uma ferramenta de recomendaçao para formaçao de grupos em ambientes virtuais de aprendizagem. *Simpósio Brasileiro de Informática na Educação*, 1657–1666. https://doi.org/10.5753/cbie.sbie.2017.1657 [GS Search].

Moore, M. G. (2013). The theory of transactional distance. In *Handbook of distance education* (pp. 66–85). Routledge. https://doi.org/10.4324/9780203803738 [GS Search].

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

Peraić, I., & Grubišić, A. (2023). Exploring student engagement in online programming courses: A two-level k-means analysis. *2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 1–6. https://doi.org/10.23919/SoftCOM58365.2023.10271619 [GS Search].

Pimentel, M. G., Ishiguro, Y., Kerimbaev, B., Abowd, G. D., & Guzdial, M. (2001). Supporting educational activities through dynamic web interfaces. *Interacting with Computers*, *13*(3), 353–374. https://doi.org/10.1016/S0953-5438(00)00042-4 [GS Search].

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451–502). Elsevier. https://doi.org/10.1016/b978-012109890-2/50043-3 [GS Search].

Puustinen, M., & Pulkkinen, L. (2001). Models of self-regulated learning: A review. *Scandinavian Journal of Educational Research - SCAND J EDUC RES*, *45*, 269–286. https://doi.org/10.1080/00313830120074206 [GS Search].

Ramos, J., Santos, L., Silva, J., & Rodrigues, R. (2020). Identificação de perfis de interação de estudantes de educação a distância por meio de técnicas de agrupamentos. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, 932–941. https://doi.org/10.5753/cbie.sbie.2020.932 [GS Search].

Razali, N. M., & Wah, Y. B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of statistical modeling and analytics*, *2*(1), 21–33. [GS Search].

Rodriguez, F., Lee, H. R., Rutherford, T., Fischer, C., Potma, E., & Warschauer, M. (2021). Using clickstream data mining techniques to understand and support first-generation college students in an online chemistry course. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 313–322. https://doi.org/10.1145/3448139.3448169 [GS Search].

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220. https://doi.org/10.22610/jevr.v3i5.60 [GS Search].

Self, J. (1990). Bypassing the intractable problem of student modelling. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems: At the crossroads of ai and education* (pp. 107–123). https://doi.org/10.1111/j.1365-2044.1983.tb14035.x [GS Search].

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India. https://doi.org/10.1007/978-1-4471-7307-6_1 [GS Search].

Urdan, T. (2010). *Statistics in Plain English* (3rd ed.). Taylor & Francis. https://doi.org/10.1111/j.1751-5823.2011.00149_21.x [GS Search].

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: A review of empirical research. *Proceedings of the tenth international conference on learning analytics & knowledge*, 524–533. https://doi.org/10.1145/3375462.3375483 [GS Search].

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, *265*(3), 66–75. https://doi.org/10.1145/2555243.2558890 [GS Search].

Zhao, X., & Okamoto, T. (2011). Adaptive multimedia content delivery for context-aware u-learning. *International Journal of Mobile Learning and Organisation*, *5*(1), 46–63. https://doi.org/10.1504/ijmlo.2011.038691 [GS Search].

Zimmerman, B., & 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. https://doi.org/10.3102/00028312023004614 [GS Search].

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*(4), 307–313. https://doi.org/10.1016/0361-476X(86)90027-5 [GS Search].

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American educational research journal*, *45*(1), 166–183. https://doi.org/10.3102/0002831207312909 [GS Search].

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 311–328). Routledge. https://doi.org/10.4324/9780203876428 [GS Search].