Exploring the Relationship between Students Engagement and Self-Regulated Learning: A Case Study using OULAD Dataset and Machine Learning Techniques

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

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

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

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

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

Abstract. *Exploring the correlation among student engagement, self-regulated learning, and academic performance through analysis of the Open University Learning Analytics Dataset (OULAD). This dataset covers course details, learner information and their interactions with the VLE. It records interactions such as resource clicks, course notes, discussions, and quizzes. Online student data was analyzed using educational data mining and three clustering algorithms: K-means, EM and Agglomerative Clustering. The results show a positive correlation between student engagement and academic performance, highlighting that greater interaction with learning resources results in better academic outcomes and shows a self-regulated approach to learning.*

1. Introduction

The increasing development of technological resources in intelligent learning environments has caused changes in the teaching and learning processes. Virtual Learning Environments (VLEs) apply different resources to support educational processes, which are often implemented to improve academic performance and students' motivation, and to reduce dropout. VLEs are online computational systems used for educational purposes in different domains and levels. The COVID-19 pandemic [World Health Organization 2020] has brought many significant changes to education around the world. Educational institutions have been forced to revamp their teaching process, using VLEs to keep the learning going while reducing the risk of exposure to the virus. With this, there has been a generation of many new data sets, which can be used to analyze the learning process of students. The datasets generated in VLEs can be used to understand how to improve learning techniques, such as self-regulated learning (SRL) [Coman et al. 2020].

Studies show that self-regulation of the learning process is related to better academic performance [Zimmerman e Martinez-Pons 1986]. In self-regulated learning, the student is the protagonist of his learning and can develop various cognitive, metacognitive, motivational, and emotional/affective strategies to self-regulate the learning process

[Lima et al. 2020]. Students with high levels of SRL skills are able to play an active role in achieving their academic goals [Pintrich e Groot 1990]. Classification of SRL profiles in online learning has mainly been based on data collected using student self-report tools [Broadbent e Fuller-Tyszkiewicz 2018] [Yot e Marcelo 2017]. Educational Data Mining (EDM) techniques can assist in measuring and profiling students more accurately than self-report tools because they use sets of real data collected from a VLE. EDM has become a viable and reliable option for detecting student behavior during their learning process. The current work investigates students' SRL profiles, using the Open University Learning Analytics Dataset (OULAD)[1]. EDM techniques were used to trace the students' SRL profiles.

We defined the research hypothesis as: Students who show higher levels of engagement and interaction with learning resources within the Virtual Learning Environment (VLE) will demonstrate more positive self-regulated learning (SRL) profiles. Furthermore, these enhanced SRL profiles will be positively correlated with better academic performance, indicating a strong connection between proactive learning behaviors, self-regulation and academic success. Thus, the study was guided by the following research questions:

RQ1: Which EDM techniques are used to identify SRL profiles in VLEs?
RQ2: Which algorithm performs better for identifying SRL profiles in the OULAD dataset?
RQ3: Which is the SRL profile of students in the OULAD dataset?
RQ4: How are students' SRL profiles correlated with students' final results?

This article has the following structure: Section 2 provides a discussion on self-regulated learning, educational data mining, and a summary of studies identifying SRL profiles, as well as presenting the dataset used in the study. Section 3 outlines the research methodology employed. Section 4 presents the results and discussions related to the research questions. Finally, Section 5 presents conclusions and final considerations.

2. Related Work

2.1. OULAD Dataset

Following a comprehensive analysis of multiple databases, including those referenced in [Costa et al. 2020], CodeBench [Lima et al. 2021], Khan Academy [Lima et al. 2021], and edX [Cobos et al. 2017], we have determined that the OULAD dataset, provided by the Open University UK, is the most appropriate for the context that we want to analyze in this work since it has more data on different types of interaction and learning resources. The types of information can be divided into three parts [Kuzilek et al. 2017]: demographic, evaluation, and VLE interaction. This dataset was preferred due to its wide use in the e-learning context, and specifically in adaptive learning contexts. The OULAD contains 7 CSV files, presented in Table 1, and requires pre-processing and transformation to extract important information in order to create forecasting models.

The use of online datasets is particularly relevant for research questions related to SRL, as it enables researchers to investigate how students regulate their learning in

[1]Dataset collected from the Open University in the UK: https://analyse.kmi.open.ac.uk/open_dataset

an online setting. For instance, the dataset includes information on students' utilization of course materials, their participation in discussion forums, and their performance on quizzes and assignments, which can be utilized to study various aspects of SRL such as goal-setting, self-monitoring, and self-evaluation. Below, the concept of SRL will be outlined.

Table 1. Tables information of the Open University Learning Analytics Dataset

Table name	Records	Description	Table attributes
studentInfo	32.593	Demographic information about the students	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result
vle	6.365	Online learning. resources and materials	id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle	1.048.574	Student interaction with the VLE resources	code_module, code_presentation, id_student, id_site, date, sum_click
courses	22	Information about the courses	code_module, code_presentation, module_presentation_length
studentRegistration	32.593	Registration of the student for a course presentation	code_module, code_presentation, id_student, date_registration, date_unregistration
assessments	196	Assessments for every course presentation	code_module, code_presentation, id_assessment, assessment_type, data, weight
studentAssessments	173.740	Assessments submitted by the students	id_assessment, id_student, score date_submitted, is_banked

2.2. Self-Regulated Learning

Self-regulated learning 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-regulation [Zimmerman e Martinez-Pons 1986].

In the review carried out by [Panadero 2017], six models of SRL were presented and compared [Zimmerman 1986] [Boekaerts 1988] [Winne e Hadwin 1998] [Pintrich e Groot 1990] [Efklides 2011] [Hadwin et al. 2011]. According to [Panadero 2017] and [Puustinen e Pulkkinen 2001], SRL models can be defined as cyclical, and they have different phases and sub-processes of self-regulation. 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.

In [Panadero 2017] the three phases are defined as: the preparatory phase comprises the analysis of tasks, the planning, the definition of objectives, and the establishment of goals; the second phase presented in the SRL models is the execution phase, where tasks are performed while monitoring progress and performance; Finally, there is the evaluation phase, where the student reflects, regulates, and adapts his learning process for future executions.

2.3. Educational Data Mining

Educational data mining (EDM) provides important information that can be used to guide students in their self-regulation of learning and to improve the effectiveness of the education system. It helps personalize the learning experience, support student decision-making and promote student success. Educational systems used for online teaching generate a large amount of data, in particular, records of student interactions with the system. These data can be used to detect interesting insights about the learning process through the use of EDM techniques. EDM is a specific area of data mining that focuses on analyzing data related to educational contexts [Costa et al. 2020].

Considering that in recent years there has been an exponential growth in the use of VLEs, both for distance learning and in support of face-to-face or hybrid education, due to the Covid-19 Pandemic, large amounts of educational data have been generated. In this context, EDM techniques can be used to analyze the educational data in these learning environments. Between the years 2000 and 2017, the main DM techniques used in the EDM process were classification and clustering [Aldowah et al. 2019]. Classification is a supervised learning technique, where a predictive model is trained from a dataset that has input and output labels. Clustering, on the other hand, is an unsupervised technique, where the dataset does not need to have labels, i.e., the output of each record is not known. Understanding SRL profiles in the EDM context is crucial, as this enables a detailed analysis of students' strategies and behaviors, generating important insights to improve the effectiveness of pedagogical practices and promote more meaningful learning outcomes.

2.4. Summary of SRL Profiles Identified From Previous Studies

Several studies found in the literature indicate that there are different SRL profiles among students in online learning environments. These profiles can be defined using EDM techniques on data collected in VLEs, through self-report or trace data. In the following, we present a summary of the works that identified the students' SRL profiles, using self-report questionnaires, data trace, and data generated in the evaluations data as sources.

In [Valle et al. 2008], authors classified students as more or less regulated, according to several indicators, through clustering. The research was carried out with higher education students. Through step-by-step linear regression analysis, it was determined which of the selected variables best predicted metacognitive self-regulation. Three significantly different self-regulated learning profiles were obtained by two-step cluster analysis with these variables. Next, ANOVA was used to analyze the relationship between SRL profiles and academic performance. This work had a significant impact by providing important insight into the relationship between self-regulated learning and academic performance, as well as demonstrating the effective application of data analysis techniques in the educational context.

The work [Barnard-Brak et al. 2020] proposed a study to examine whether there are profiles for self-regulated learning skills and strategies among students. They performed two studies with two different samples. The Online Self-Regulated Learning Questionnaire (OLSQ) was applied. They used latent class analysis to identify SRL profiles, resulting in the presence of five distinct self-regulated learning profiles replicated in both study samples: super self-regulators, competent self-regulators, premeditated self-regulators, performance/reflection self-regulators, and no or minimal self-regulators. The study [Li et al. 2018] analyzed student tracking data on the VLE. They used records related to access to learning materials, completion of questionnaires, and response records to profile the SRL. The K-means clustering algorithm was applied and four distinct groups were identified: 1)early graduates, 2)late graduates, 3)early dropouts, and 4)late dropouts.

A mixed approach was used in the work of [Ainscough et al. 2019]. Trace and self-report data were used to define the SRL profiles. They were divided into three groups: high self-regulators, medium self-regulators, and low self-regulators. A two-step cluster analysis was used to group students. The first step was the formation of the pre-cluster. In the second step, the hierarchical clustering algorithm was used to merge the pre-clusters, leading to the three different clusters. The authors [Costa et al. 2020] analyzed data from a Ubiquitous Educational Environment using data clustering techniques to observe student behavior in learning sessions. The authors applied the K-means algorithm to perform data clustering. Two groups were found, and one of these groups showed strong evidence of students' self-regulation capabilities. The review presented by [ElSayed et al. 2019] showed that there is a lack of studies to define which EDM algorithm has a better performance in identifying SRL profiles through tracking data collected in VLEs.

3. Methodology

The study followed a five-stage methodology: data extraction from the OULAD dataset, selection of relevant files, data preprocessing, application of clustering algorithms, and analysis of results to identify student SRL profiles. Most relevant aspects are following described.

3.1. OULAD

The OULAD dataset was collected during the years of 2013 and 2014 from Open University Uk. OULAD contains data about courses, students, and their interactions with Virtual Learning Environment (VLE) for seven courses [Kuzilek et al. 2017]. While the dataset also includes demographic data, the current study focuses on the aggregated clickstream data from student interactions within the VLE. The study specifically utilized tables extracted from *studentInfo*, *studentVle*, and *vle* files. Table 1 shows the information contained in each file [Araka et al. 2022]. Interactions contained in the dataset reflect the number of clicks to various resources and learning activities, such as course notes in HTML or pdf format, as well as learning activities that involve discussion forums and quizzes.

3.2. Data Preparation

In this subsection, we present the details of the data pre-processing. After feature extraction, the resulting table was summarized in Table 2. We used the three files mentioned in Table 1, and extracted the data using the unique identification of each student in the

entire database, i.e., *id_student*, as the key. The pre-processing of the data and the construction of the final dataset were performed using the Pandas library in Python, which offers user-friendly data structures and data analysis tools for handling and manipulating large datasets [McKinney et al. 2010].

Table 2. Summary of the dataset after feature engineering

Category	N. Attributes	Attributes	Type
Sum click for each VLE activity_type	20	Sum_Clicks_{resources, oucontent, url, homepage, subpage, glossary, forumng, oucollaborate, dataplus, quiz, ouelluminate, sharedsubpage, questionnaire, page, externalquiz, ouwiki, dualpane, repeatactivity, folder, htmlactivity}	Numeric

In order to reduce the complexity of the data, the attributes were aggregated. Similar attributes were grouped together, resulting in the reduction of the original dataset from 20 attributes to 5 numerical attributes, as described in Table 3. The attributes *folder*, *sharedsubpage*, and *repeatactivity* were excluded from the dataset due to their low frequency and lack of significance for the current study's context. This resulted in a final dataset with 29741 rows (students) and 5 columns (attributes).

Table 3. Aggregation of attributes

Activity type	Grouping
collaborative	forumng, outcollaborate and ouelluminate
activities	quiz and questionnarie
access	homepage, resource, ouwiki, page, and htmlactivity
resource_ext	externalquiz and url
views	glossary, outcontent, dualpane, subpage, and dataplus

At this stage, several statistical analyses were conducted. Firstly, the Anderson-Darling test [Anderson e Darling 1952] was performed in order to determine if the data followed a normal distribution. The normal distribution describes a symmetrical distribution of data points around a central mean, where most of the data falls near the mean and gradually tapers off towards the tails. When the dataset does not conform to this assumption, it can have important implications for the validity and interpretation of analyses. Based on the test results, it was determined that the data did not follow a normal distribution, leading to the use of non-parametric variance tests. In order to identify how one variable behaves when another variable is varying, a correlation matrix of the variables was constructed using the non-parametric Spearman's correlation coefficient. This coefficient measures the monotonic relationship between two variables [Spearman 1961].

3.3. Data Clustering

This article investigates the performance of three Clustering algorithms to track the student's SRL profile, considering clicks on the resources available in the VLE: K-means, Expectation Maximization (EM), and Agglomerative Clustering. The choice of clustering algorithms for our study was based on the nature of our data and objectives. We chose from different categories of algorithms: partitional, hierarchical, and model-based. The algorithms were applied with resources from scikit-learn libraries[2].

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

K-Means is a partitioning algorithm that divides a set of X of n samples into K disjoint groups, each described by an average u of the samples in the group. This mean is called the centroid of the group [Hastie et al. 2009]. Agglomerative Clustering a hierarchical algorithm and uses a bottom-up approach to perform the clustering, that is, each element of the dataset starts in a group and, at each step, the pairs of elements merge according to their proximity [Jain e Dubes 1988]. EM is designed to estimate the maximum likelihood parameters of a statistical model in many situations, such as the ones where the equations cannot be solved directly. It is an iterative technique consisting of two main steps: the E step (Expectation) and the M step (Maximization) [Moon 1996].

4. Results and discussion

In this section, we examine the experimental results obtained from the three clustering algorithms. First, we present and evaluate the outcomes of the three clustering algorithms, aiming to determine the most suitable algorithm with the optimal number of clusters. The Silhouette Coefficient, Dunn's Index, Calinski-Harabasz, and Davies-Bouldin were used as measures.

The Silhouette Coefficient is a metric that measures how well each data point fits into its assigned cluster based on the distance between the data point and other points within its cluster, as well as the distance between the data point and the points in other clusters [Rousseeuw 1987]. The metric was chosen because it performs a holistic assessment of cluster quality, considering both internal cohesion and separation from neighboring clusters. Dunn's Index measures the distance between the closest clusters relative to the average size of the clusters [Jain e Dubes 1988], used to identify optimal cluster compression while maintaining proper separation between clusters. Calinski-Harabasz is a measure of the density and separability between groups, used to pinpoint well-defined and densely-packed clusters with clear separations, while Davies-Bouldin measures the similarity between the group and its closest group [Furlanetto et al. 2022], the metric was selected for its ability to highlight distinct and well-separated clusters.

Table 4 presents the cluster validation measures for the different clustering algorithms used in this study. The results show that Agglomerative Clustering had the best performance in terms of the Silhouette Coefficient (0.65) and Dunn Index (0.0038). On the other hand, K-means had the best performance in terms of Calinski-Harabasz (20419) and Davies-Bouldin (0.936) measures. Therefore, further analysis of the grouping data is necessary to determine the algorithm with the best performance. We performed a descriptive analysis of the clusters, as shown in Table 5, which provides the mean and standard deviation of each attribute. After analyzing the clusters generated by each algorithm, we chose to use K-means because the sizes of the generated clusters were more balanced than those obtained with Agglomerative. Uneven cluster sizes can indicate that the data is not well-clustered. Balanced clusters are important as they ensure that each cluster represents a complete subset of the data, decreasing the risk of skewed insights. This balance increases the reliability and generalizability of our results, ensuring that no subset is overemphasized.

Figure 1 highlight the mean values of each attribute in the clusters generated using the K-means algorithm. To determine if there were statistically significant differences between the means of the attributes, a non-parametric Kruskal-Wallis test was conducted,

Table 4. Optimal algorithm and cluster evaluation results

		Clusters			
Algorithm	**Validation measure**	**2**	**3**	**4**	**5**
K-means	Silhouette Coefficient	0.64	0.62	0.52	0.54
	Dunn Index	0.0021	0.0025	0.0018	0.0015
	Calinski-Harabasz	**20419**	17165	16465	15101
	Davies-Bouldin	**0.936**	0.954	1.01	1.002
Agglomerative Clustering	Silhouette Coefficient	**0.65**	0.64	0.44	0.45
	Dunn Index	**0.0038**	0.0035	0.0012	0.0012
	Calinski-Harabasz	18170	14837	13929	12571
	Davies-Bouldin	0.945	0.878	1.112	1.145
Expectation Maximization	Silhouette Coefficient	0.36	0.21	0.10	0.09
	Dunn Index	0.0007	0.0003	0.00016	0.00011
	Calinski-Harabasz	10040	8382	6433	5840
	Davies-Bouldin	1.120	1.376	1.500	1.660

Table 5. Descriptive statistics per cluster using agglomerative and k-means

	Agglomerative		K-means	
Attribute	**Cluster 0 (N=29685)**	**Cluster 1 (N=56)**	**Cluster 0 (N=25165)**	**Cluster 1 (N=4576)**
collaborative	311.24 ± 576.52	8442.64 ± 1576.74	191.26 ± 295.85	1070.57 ± 1363.12
activities	298.20 ± 555.62	884.71 ± 1240.22	160.36 ± 273.14	1063.38 ± 961.80
access	367.86 ± 447.73	3760.67 ± 2294.64	248.17 ± 237.33	1067.59 ± 791.44
resource_ext	25.72 ± 43.23	179.30 ± 165.51	18.62 ± 237.33	66.66 ± 77.52
views	592.34 ± 844.06	1744.51 ± 1875.97	322.12 ± 342.17	2092.45 ± 1176.59

as the data did not follow a normal distribution, which was confirmed by the Anderson-Darling test. The Kruskal-Wallis test confirmed the statistical significance of the differences between the clusters. The clusters depicted in Figure 1 reveal the discrepancies between the resources utilized by the groups. Specifically, Cluster 1 exhibits higher mean values across all attributes, with *"views"* and *"access"* standing out. This suggests that the students in Cluster 1 were more actively involved in utilizing the resources available on the VLE. This finding is consistent with [Zimmerman e Martinez-Pons 1986], which indicates that self-regulated learning involves proactive students who take responsibility for their own learning process. Moreover, students who engage in more interactions on the educational platform tend to achieve better academic outcomes.

One can also highlight that collaborative resources, such as forums, had a higher average in cluster 1. These resources are important for fostering student-teacher and student-student interaction. By using these resources, students can engage in discussions, which can help them self-monitor and define strategies for performing tasks [Kitsantas 2013]. According to [Zimmerman e Martinez-Pons 1986], accessing information through external URLs is considered a self-regulated learning strategy (SRL). Therefore, it can be inferred that students who exhibit more interactions with external resources are more likely to have self-regulated behavior. Interactions (*"activities"*) were grouped with clicks in quizzes and questionnaires. Students with self-regulation characteristics tend to self-evaluate more often [Kitsantas 2013]. Which again indicates evidence of student self-regulated behavior in this analysis.

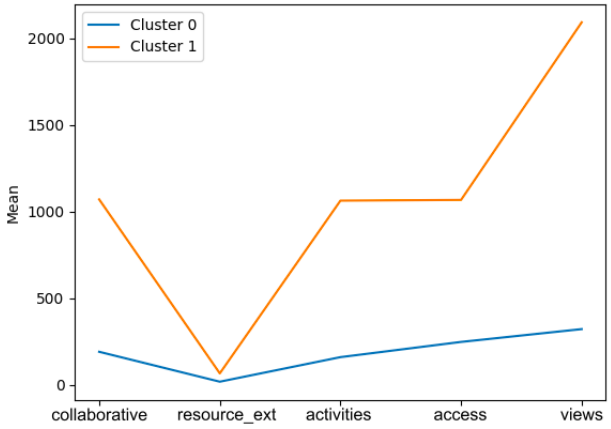

Figure 1. K-means mean in each attribute

Finally, correlation analysis was used to identify the association between the SRL profiles and students' academic performance. The chi-square test was carried out to establish the correlation between the SRL profiles formed by K-means grouping and students' final results. After calculating the ($p\_value = 0.00 \leq 0.05$), we can conclude that there is a significant relationship between the SRL profiles and the students' final results. Figure 2 shows the final average of approved and failed in each cluster. Therefore, we can conclude that students with more interactions show a self-regulated profile that tends to produce better performance results.

The figures in this study provide valuable insights into the relationship between student engagement and academic performance. Specifically, Figure 1 illustrates the differences in resource utilization between the two clusters. Cluster 1, which exhibited higher mean values across all attributes, suggests that students in this cluster were more actively involved in utilizing the resources available on the VLE. Moreover, the findings presented in Figure 2 suggest that there is a positive correlation between student engagement and academic performance. Students who had higher levels of interaction with the learning resources exhibited a self-regulated learning profile that was associated with better academic outcomes. The higher proportion of successful students in Cluster 1, as compared to Cluster 0, suggests that greater engagement with the VLE resources may have contributed to their academic success. Therefore, it can be concluded that student engagement is an important factor in promoting self-regulated learning and achieving academic success. Next section presents conclusions and future work.

5. Conclusion

In this work, we define the main EDM techniques used to identify SRL profiles of students in VLEs. We present the main clustering algorithms employed in the literature and the primary sources of data: self-report and trace data. Three data mining algorithms from different categories (partitional, hierarchical, and model-based) were applied to the dataset produced with interactions collected from OULAD. The performance of the clustering algorithms was evaluated using internal validation measures to determine the most effective algorithm and the optimal number of clusters. Based on our dataset, the K-means algorithm with a cluster size of 2 ($K = 2$) produced the most favorable result in terms of clustering, considering both validation measures and cluster formation. When examining

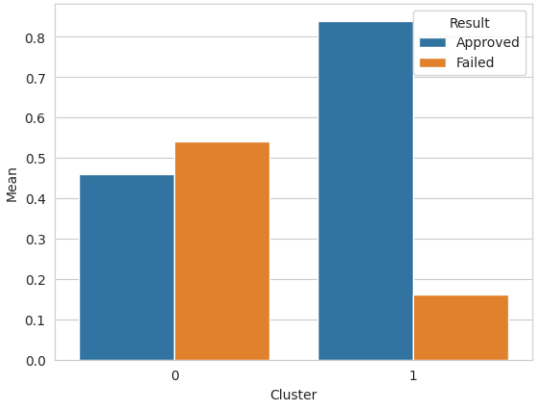

Figure 2. Relationship Profile SRL based on Final Results

the resulting clusters, it can be highlighted that there is evidence of two distinct profiles of self-regulated learning (SRL) present in the dataset. By analyzing student behaviors inferred from the OULAD dataset, we mapped the two groups into two SRL profiles: Cluster 0 (No self-regulation) and Cluster 1 (Evidence of self-regulation). Students' final results were accessible via OULAD, which enabled us to examine the correlation between the grades and clustering. The analysis revealed that students who utilized more self-regulated learning (SRL) strategies, as indicated by their level of interaction with the system, tended to achieve better performance outcomes, as demonstrated by the results in Cluster 1. On average, the pass rate for students in Cluster 1 was (83%), while in Cluster 0, it was only (46%).

An important finding from this study is that grouping students based on their self-regulated learning profile can offer a valuable approach to comprehending student behavior in online environments. This classification enables educators and tutors to provide targeted assistance and guidance to students based on the specific requirements of their SRL group. We would like to highlight some limitations of this work. First, the relationship we made between students SRL profiles and their final result was based on a correlation analysis, the results may therefore not have exposed all the factors that could contribute to their approved or failed. So, for future work, it is interesting to consider other variables besides the number of clicks on the resources of the virtual learning environment.

References

Ainscough, L., Leung, R., Colthorpe, K., e Langfield, T. (2019). Characterizing university students' self-regulated learning behavior using dispositional learning analytics. In *HEAD'19. 5th International Conference on Higher Education Advances*, páginas 233–241. Editorial Universitat Politècnica de València.

Aldowah, H., Al-Samarraie, H., e 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.

Anderson, T. W. e Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, páginas 193–212.

Araka, E., Oboko, R., Maina, E., e Gitonga, R. (2022). Using educational data mining techniques to identify profiles in self-regulated learning: an empirical evaluation. *The International Review of Research in Open and Distributed Learning*, 23(1):131–162.

Barnard-Brak, L., Paton, V. O., e Lan, W. Y. (2020). Profiles in self-regulated learning in the online learning environment. *International Review of Research in Open and Distributed Learning*, 11(1):61–80.

Boekaerts, M. (1988). Motivated learning: bias in appraisals. *IJER*, 12(3):267–280.

Broadbent, J. e Fuller-Tyszkiewicz, M. (2018). Profiles in self-regulated learning and their correlates for online and blended learning students. *Educational Technology Research and Development*, 66.

Cobos, R., Wilde, A., e Zaluska, E. (2017). Predicting attrition from massive open online courses in futurelearn and edx. In *Workshop at the 7th International Learning Analytics and Knowledge Conference*, Vancouver, Canada.

Coman, C., Țîru, L. G., Meseșan-Schmitz, L., Stanciu, C., e Bularca, M. C. (2020). Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability*, 12(24):10367.

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

Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist*, 46(1):6–25.

ElSayed, A. A., Caeiro-Rodríguez, M., MikicFonte, F. A., e Llamas-Nistal, M. (2019). Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review. In *World conference on mobile and contextual learning*, páginas 46–53.

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

Hadwin, A. F., Järvelä, S., e Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning. In Zimmerman, B. J. e Schunk, D. H., editors, *Handbook of self-regulation of learning and performance*, Educational psychology handbook series, páginas 65–84. Routledge/Taylor & Francis Group.

Hastie, T., Tibshirani, R., Friedman, J. H., e Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Jain, A. K. e Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Kitsantas, A. (2013). Fostering college students' self-regulated learning with learning technologies. *Hellenic Journal of Psychology*, 10(3):235–252.

Kuzilek, J., Hlosta, M., e Zdráhal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, 4:170171.

Li, H., Flanagan, B., Konomi, S., e Ogata, H. (2018). Measuring behaviors and identifying indicators of self-regulation in computer-assisted language learning courses. *Research and Practice in Technology Enhanced Learning*, 13:1–12.

Lima, G., Araújo, R., e Dorça, F. (2020). Uma análise dos recursos tecnológicos utilizados na estimulação da aprendizagem autorregulada em ambientes educacionais na Última

década. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, páginas 732–741, Porto Alegre, RS, Brasil. SBC.

Lima, M., Carvalho, L., Oliveira, E., Oliveira, D., e Pereira, F. (2021). Uso de atributos de código para classificação da facilidade de questões de codificação. In *Anais do Simpósio Brasileiro de Educação em Computação*, páginas 113–122, 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, páginas 51–56. Austin, TX.

Moon, T. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60.

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

Pintrich, P. R. e Groot, E. V. D. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, páginas 33–40.

Puustinen, M. e Pulkkinen, L. (2001). Models of self-regulated learning: A review. *Scandinavian Journal of Educational Research*, 45:269–286.

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.

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

Valle, A., Núñez, J. C., Cabanach, R. G., González-Pienda, J. A., Rodríguez, S., Rosário, P., Cerezo, R., e Muñoz-Cadavid, M. A. (2008). Self-regulated profiles and academic achievement. *Psicothema*, 20(4):724–731.

Winne, P. H. e Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. In Hacker, D., Dunlosky, J., e Graesser, A., editors, *Metacognition in Educational Theory and Practice*, páginas 277–304. Erlbaum, Mahwah, NJ.

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.

Yot, C. e Marcelo, C. (2017). University students' self-regulated learning using digital technologies. *International Journal of Educational Technology in Higher Education*, 14.

Zimmerman, B. e 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. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, 11(4):307–313.