Data Analysis of Short - term and Long - term Online Activities in LMS

Carmen Carrión

School of Computer Engineering, University of Castilla-La Mancha, 02071 Albacete, Spain

Abstract **– Online teaching activities based on increasingly used computer-based educational systems lacks standard rules for its implementation. This paper describes the design of online training activities using Moodle as a Learning Management System (LMS) and, evaluate short-term and long-term students' learning outcomes applying data mining techniques. Clustering and classification algorithms are combined to uncover valuable, non-obvious students' patterns from a well-defined collection of data. Data results from online quiz-based activities in a subject of Computer Science show that students who are not engaged in the training activity during the short-term learning process fail. Data analysis also shows that the number of trials is a key attribute. Hence, it is important to develop user-friendly online activities with real-time feedback based on student behaviour. Moreover, according to our experiment, online training activities decrease in efficiency over time.**

Keywords **– higher education, Learning Management System, Data Mining, students' behaviour, students' outcomes.**

1. Introduction

The COVID-19 pandemic stresses the need for innovative learning technologies based on online activities [1].

DOI: 10.18421/TEM112-01
https://doi.org/10.18421/TEM112-01

Corresponding author: Carmen Carrión,
School of Computer Engineering, University of Castilla-La Mancha 02071-Albacete, SPAIN.
Email: carmen.carrion@uclm.es

Received: 27 January 2022.
Revised: 15 March 2022.
Accepted: 23 March 2022.
Published: 27 May 2022.

The spread of COVID-19 and its variants means that many educational centres have to switch to an educational model in which part, if not all, of their teaching is online. Thus, the teacher incorporates open educational resources or activities based on Learning Management Systems (LMS) to complement face-to-face classes [2], [3], [4]. Massive Open Online Courses (MOOCs) even appear as an extreme expression in which all teaching is online [5]. However, no standard guides exist on how to implement an online activity in a course or grade, being always left to the best judgment of the teacher in charge [6], [7].

In any online educational activity, which is to be developed, computer-based educational systems become a fundamental pillar. So, LMSs, such as Moodle, Edmodo or Canvas are becoming widely used in higher education [4]. Nowadays, LMSs are mainly used as a static information exchange tool [8]. On the one hand, the teacher uploads information about the syllabus and slides of the course and, on the other hand the students upload the results of the laboratory practices and exercises sent as homework [9]. But, LMSs often have configurable functional features, such as quizzes, puzzles or blogs that allow dynamic interaction and students' collaboration to be included in the online teaching-learning process. The primary purpose for the use of these functional features is to increase students' motivation, engagement and academic results. Notice that in online learning environments measuring the efficiency of the developed activities, as well as, adapting the teaching process according to the feedback received from the students is not straightforward as it is the case in traditional face-to-face classes. Efficiency analysis of online activities is only possible if sufficient information about the individual characteristics of the students is obtained through interactions with the environment. Researchers usually apply different methodologies to the educational context in order to discover hidden knowledge and patterns in the learning process [10]. Out of all, Data Mining (DM) techniques have become increasingly prominent for improving teaching-learning process. They make use of mathematical algorithms to extract non-obvious behavioural patterns of interest from a large data set [11].

In this paper, we emphasize on understanding the education process in its full complexity, leveraging teacher judgment, and applying DM techniques to gain insight into the students' learning process [12], [13], [14]. This task is not easy, and online behaviour of students and their predicted performance are in contrast between the studies carried out up to now [13], [16], [17].

To tackle this issue, more research is required, and this paper takes a step forward by establishing a set of objectively verifiable indicators for online teaching-learning activities and, examining measures related to short-term and long-term knowledge.

Hence, the major aim of this paper is to extract valuable information about the achievement of online activities in the short-term and long-term in order to guide teachers in their development and to make a continuous improvement in learning outcomes.

Data-driven analysis collects student behaviour during the completion of online activities via LMS and makes use of DM clustering and classification techniques. The paper presents a use case for Computer Engineering, but the study can be applied to different educational environments, courses and subjects. Specifically, the aim of this paper is to answer questions like these:

- Is there a relationship between students' online activity behaviours and their short-term and long-term outcomes?
- Can relevant information be extracted from the online activity interaction data to understand its impact, guide the teacher, and eliminate, if any, activity-related learning deficiencies?

Education is alive process, and the proposed data-drive analysis of the online activities will support the teacher with the feedback needed to succeed in the teaching-learning process. To sum up, the key contributions of this paper are:

- The proposal of a generic and reproducible online teaching-learning activity is based on opensource tools.
- The definition of a set of quantitative parameters is useful for data analysing online activities.
- The use of DM as a tool is to discover significant patterns to model students' trends both in the short-term and long-term learning process.
- The analysis and evaluation are of a practical use in computer science.

The paper is organized as follows. In Section 2, we present the results obtained and techniques employed by the related work. Details of our life cycle for designing and analysing online activities are discussed in Section 3. Then, Sections 4 and 5 detail the data parameters and data mining analysis applied in our approach, respectively. After that, a use case in Computer Science is presented in Section 6, followed by the data analysis results in Section 6-a. Finally, Section 7 presents the conclusions and the future work.

2. Related Work

Different DM techniques for finding hidden knowledge and patterns in the learning process are applied to educational environments [10]. Several systematic literature reviews have been conducted in education to describe the state-of-the-art in this area [18], [19], [20]. Clustering and classification DM techniques provide good results in predicting the results [21].

In [18], authors present over three decade's systematic literature review on clustering algorithms and its applicability in EDM. The paper outlines that educational data are non-independent in nature and clustering can provide a relatively unambiguous outline of student learning style as a function of several variables. In [19], authors show that most of the researching work in the area of predicting students' performance is looking at predicting attainable academic metrics such as exam grade, course grade, program retention or dropout or assignment performance [22], [23]. For example, in [23] authors use decision trees to apply classification techniques and detect factors linked to academic performance in large-scale assessments. Their results indicate that personal factors are the most indicative for academic performance, followed by school-related and social factors. Moreover, according to [19], data of students' activity is one of the least explored to predict students' performance, that is the focus of this paper.

The effect of students' online interaction with Moodle and their relationship with achievement is examined in [15]. The variables considered in [15] are time task, time theory, time forums, word forums, relevant actions and procrastination. Additionally, final marks were extracted from the performance of the subject. Using k-means clustering techniques they found that more activity in the LMS does not assure better results. Moreover, they got that the students who hand in the task later were more likely to receive a lower score, that is, the more procrastination, the worse the performance.

In [16] significant indicators from the LMS data such as regular study, total viewing time, sessions, late submissions, proof of reading the course information packets, and messages created were chosen as predictors. The results revealed that regular study was the strongest predictor of course achievement, followed by late submissions, sessions, and proof of reading the course information packets. However, students' total viewing time and messages created were not significant.

In [13], authors use time spent for accessing online learning materials (video, slides, etc.) to classify students into different clusters. They found that behaviours of accessing online learning materials were associated with learning performance. More precisely, the students who invested more time and effort in viewing the online learning materials had better learning performance.

In [17], authors use k-mean clustering algorithm to examine if students' interaction with different online learning activities affects students' learning performances, motivation, and self-regulated learning strategies. Their results conclude that the students who spend more time in the learning activities got higher academic outcomes.

Nevertheless, at this point we have to highlight that the conclusions drawn in [15], [16] contradict the abovementioned results in [13], [17]. These facts may be due to the analysis were done under different contexts, type of courses and students backgrounds [24], [17]. In any case, these facts underline the need for further studies to establish some baselines and avoid the contradictions, as well as, to set up a set of objectively verifiable indicators for the teaching-learning process. For this reason, this paper presents the results obtained for a use case using a generic methodology based on the analysis of online activity data for the short-term and long-term.

3. Development and Evaluation of Short-term and Long-term Online Activities

The methodology applied in this paper to model and analyze the effectiveness of online activities using Moodle LMS includes a complete life cycle that combines two main roles: the in-class and the off-stage. Figure 1 details the steps involved in the development of the online activities. Next, we will explain all the involved steps in detail.

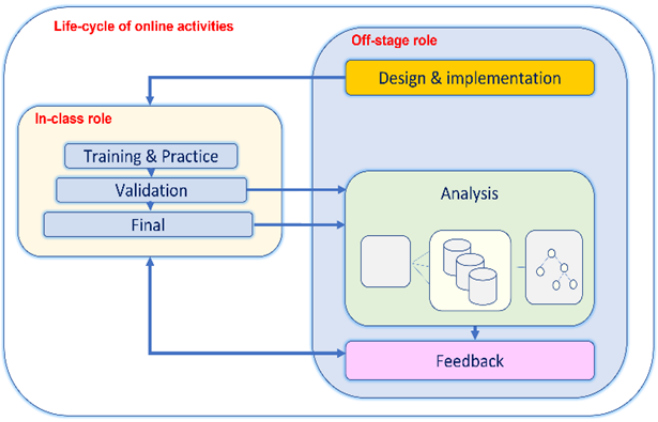

Figure 1. Proposed methodology to develop online activities

The in-class role comprises those stages in which the student participates actively even though they are conducted online. More precisely, the in-class role includes the training, the validation, and the final step.

- Training and Practice step: Online training activities are web-based activities accessible via the Internet that are available 24/7 up to a prefixed day on the digital platform. So, students can access the online training activity whenever they want, use the time they need and repeat the online training activity as many times as they wish. As a counterpart to face-to-face training activities lack the direct support of the teacher, where the teacher provides the appropriate coaching and supervision while the activity is running in the classroom. Hence, it is essential to give specific and easy-to-understand instructions on how to carry out the activity.
- Validation step: After the training step the students should have acquired new skills and knowledge. Therefore, the validation step allows them to check their improvement. At this point, the student receives the first quantifiable performance indicators about his or her progress. This step can be an important turning point for students, being a very stimulating phase in the learning process or just the opposite. The parameters collected in this step are related to short-term learning. In this paper, this term will be defined as the learning process over the course of a single session or during the training period of a single online activity.
- Final step: This is the last step of the in-class role and provides information about the long-term learning process. Data collection includes both cognitive and affective outcomes.

The students will make theoretical questionnaires and practices exercises at the end of the course for measuring the cognitive outcome and will complete post-test questionnaires for affective outcomes.

The off-stage role is characterized by the fact that the student does not actively participate in it, although it is crucial for the achievement of the objectives for which the online activity is carried out. The off-stage role includes the design and implementation, the analysis and the feedback step.

- Design and implementation step: Learning activities are created and designed to achieve learning objectives. So, in this step the teacher addresses the challenges of analyzing the capacity and feasibility of the online activities to work on the learning outcomes: the complexity, the context and the timeline of the activity have to be taken into account. It is compulsory to examine carefully the previous student's understanding on the topic, as well as, the activity complexity to reduce the negative outcome of poor performance. Then, a well-structured support has to be developed before starting the learning activity. Both the required theoretical knowledge and a detailed description of how to do the activity have to be provided to the student in order to successfully achieve the goals of the activity. Moreover, an updated scheduling timetable has to be

presented to the students in order to know in which sessions the activity will be running and also some important deadlines. For example, if the activity requires the production of a deliverable, the deliverable have to always be uploaded to the LMS on time. Setting deadlines for activities in advance is especially important if they involve extra work outside of class time. To sum up, the design and implementation step is a key phase in the off-stage role that includes:

- The teacher's training period in which the teacher searches for new teaching approaches and learns the details of using new digital tools.
- The design period in which the teacher plans the activities considering the learning objectives and the time available for the activity. The teacher also has to design the observable parameters for further analysis. A detailed description will be provided in subsection 4.
- Finally, in the get-ready step the teacher prepares all the materials. This step may involve for example the access to a digital platform, prepare some hints and prizes and/or design a multiple-choice test.
- Analysis step: Both a statistical exploration of data and DM techniques can be used to discover useful information which will allow us to improve the scheduling and the design of the activity. The first steps of the DM analysis include data collection and pre-processing. In some cases, data collection for students requires knowledge of applicable personal privacy laws and regulations. In the proposed data analysis, just after the pre-processing, a clustering technique is combined with classification techniques to observe students' patterns and accurately predict important features. In short, data is transformed into useful information for decision-making. More details of the proposed data analysis are provided in Section 5.
- Feedback step: Based on the reports obtained in the previous step, we will detect and even anticipate possible drifts in the teaching-learning process. It is about predicting deviations and non-effective performance in the classroom, as well as decreasing the dropout rate. The effectiveness quality of the activity is improved by applying successive refinements that consider the feedback obtained from the data analysis. The key idea is to adapt the teaching-learning activity solving the detected problems and reinforcing the mechanisms that do work. It is important to note that the feedback step may not be the last one of the season. Data analysis can trigger a reactive mechanism on a continuous basis while the activity is in progress. For example, it would be possible to collect and analyze information during the training step.

4. Collecting Parameters for Data Mining

Improving the efficiency of teaching activities tackles the definition of suitable metrics. It is necessary to collect some quantifiable parameters to measure both the students' behavior and performance that denote the level of success. Examples of these parameters may be the percentage of tasks completed or the score obtained in a test. Specifically, the parameters employed in this research have been classified considering their target end-goal as cognitive and behavioral. Cognitive parameters are directly related to academic scores or grades and, behavioral parameters are related to the student's commitment and attitude during the course. Among the parameters used to analyze the student behavior we have collected:

The Trial **(T)** is measured as the number of times as student completes an activity. Note that some online quizzes can be done as many times as students need but some just-in-time teaching activities can be done only once.

The Reinforcement **(Re)** computes the number of trials done by the student after having obtained the maximum score in the activity.

The Readiness **(R)** is a measure of the student's willingness to do. More precisely, in this work readiness has been calculated as the time the student takes the task in-advance of the delivery time.

The learning Velocity **(V)** is measured as the number of trials done by the student to obtain the maximum mark. The parameter reflects in some way the learning capacity of the student. It can be a useful metric to give personal assistance to the student.

The Cost **(C)** is measured as the time employed for each student to complete the learning activity.

The Participation **(P)** is a global metric that measures the total number of students that participate in a teaching-learning activity.

To measure the cognitive outcomes, we have collected:

The Initial Score **(IS)** provides a quantitative measure of the student's initial level of knowledge on the topic. A high score denotes a high probability that the student already had the targeted skills before starting the activity while a low value indicates the opposite.

The Recent Score **(RS)** measures the short-term knowledge acquired by the student. RS takes into account what students know at the end of a training and practice step without considering the mistakes made during that learning process. This metric should be collected as soon as the practice step concludes.

The Final Score **(FS)** computes the long-term knowledge. The key idea is to tackle the final activity, probably at the end of the course, to measure

the knowledge acquired by the students thanks to the developed activity.

The Effectiveness **(E)** calculates the number of students that did not possess the targeted knowledge before the activity but ended up successfully at the end.

5. DM in the Off-stage Role

In this section, we outline the steps to be carried out to extract value from the data obtained from the students' interaction with the Moodle activities. We want to verify whether the online activities have been designed successfully and therefore have boosted the teaching-learning process. In general, the learning data mining process consists of the following steps: collection, pre-processing, mining, and evaluation. Figure 2 outlines the main steps of our approach.

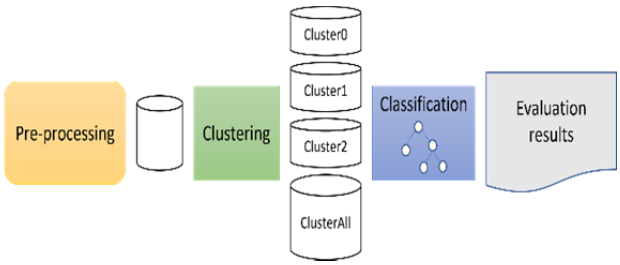

Figure 2. Workflow for data analyzing the online activities

Moodle provides us with information about the materials accessed by the students and records all the clicks made by the students when they navigate through the different resources. This data constitutes the student's digital footprint in Moodle and can be exported in different formats to start processing. We have specifically exported the information provided by Moodle 3.5v Institutional [4]. By processing these digital traces we will get the parameters indicated in Section 4.

More precisely, from all the available data, we have extracted the data collected by the Moodle quiz module for each online activity and the final grades obtained by the students. It should be noted that the quiz module collects information about the student's name and the way of accessing the platform. It also includes the access and completion times of the quiz and the score obtained. These data are recorded each time the student takes the activity and in our case there may be several trials. The analysis of the data will allow us to make a short-term cognitive analysis while the final results will allow us to analyze the long-term cognitive results. Moodle provides this data in text files with *csv* format.

In the data pre-processing stage, the data are purged. First of all, duplicate attributes and unnecessary attributes such as e-mail or access IP address are removed. Data associated with teachers are also removed and, all data are filtered to consider only those online activities that have been successfully completed. In addition, at this stage, the students' first and last name is converted into a numeric fingerprint that uniquely and anonymously identifies each student.

After that, using a computer spreadsheet application, simple calculations were performed to obtain the quantitative values of the metrics defined in the Section 4. In a course, as many short-term online activities (A_i) can be performed as many learning objectives are defined (N). Therefore, for each metric defined in Section 4 we will obtain a vector of N values. In our analysis, we have computed a global value for each parameter to avoid non-significant fluctuations. In general, the global value of a parameter (labeled as K) is computed as a weighted (Wi) average of each individual activity (Ai) according to Eq. 1:

$$K = \frac{\sum_{i=1}^{N} A_i * W_i}{\sum_{i=1}^{N} W_i} \quad (1)$$

Lastly, in the pre-processing step the numerical values are transformed into discrete attributes following unsupervised methods. These methods allow the discretization of continuous values into categorical classes following different criteria (i.e. equal width, frequency). In our case the parameters defined to analyze the behavior of the students have been discretized following the equal width method with four intervals. However, for the discretization of the cognitive values a manual criterion has been defined. It has four intervals and the labels Fail, Pass, Good and Excellent. Thus, for the cognitive metrics, a value greater than or equal to 9 will be classified as Excellent and a value less than 5 will be classified as Fail. In addition, the label Pass will be assigned for values greater than or equal to 50 but less than 70 and, label Good for values greater than or equal to 70 but less than 90.

After preprocessing, clustering and classification techniques are performed. Our approach follows the technique detailed in [25]. In general, a clustering algorithm will group students with similar properties. Thus, in this study we will make use of two clustering algorithms, one manual and one automatic. The manual clustering groups students according to the students' final marks while the automatic clustering uses a clustering algorithm based on all the behavioral data. In this case we will use the Expectation-Maximization (EM) clustering algorithm that will group students with similar characteristics without the need to indicate the number of clusters. More precisely, the EM algorithm finds maximum likelihood estimators of parameters in probabilistic models that rely on unobservable variables.

Then, for the purpose of observing students' patterns we apply some classification techniques. The idea is to use the criterion variable (RS or FS) as a reference for the algorithm to split the input data set into mutually exclusive subgroups. The aim of using the classification algorithm is to model and predict IR and IF for each case of the input dataset. Note that RS is the criterion variable for the short-term teaching-learning process while FS is used for the long-term process. The better the students are ranked on the criterion variable, the higher the validity of the model. Individual well-known classification techniques, such as C4.5, LMT, *RandomTree* or *HoeffdingTree* should be explored to increase the final accuracy and precision of the system.

6. On-line Activities Carried out in the Subject of Computer Architecture

The online activities outlined in Section 3 have been implemented this academic year in the third year subject of Computer Engineering called Computer Architecture at the University of Castilla La Mancha (UCLM). The university provides an institutional Moodle, currently in its version 3.5v [26].

In particular, eight short-term online activities were scheduled, one for each learning objective or subject topic. Participation in these online activities was not compulsory for students, but participation was encouraged using credits. So, students involved in the online activities earn credits that are redeemable for a percentage of the final course grade, in our case, up to 10% of the final grade of the course. It should be noted that all students enrolled in the course participated in the online activities.

In the design and implementation step of the off-stage role, a bank of multiple-choice questions classified by topic is generated. In our case we have 8 topics and 542 questions. Then, the Moodle quiz module is used to generate the eight online activities and schedule them appropriately over time throughout the term. Each of these quizzes is available for the *Training&Practice* step and the student can repeat the quiz as many times as he/she wishes. A quiz will consist of 10 multiple-choice questions randomly selected from the question bank within the topic to be studied. Once the training period is over, the validation stage takes place and the students perform the final test of the short-term online activity.

In addition, a final global questionnaire is carried out at the end of the course. This questionnaire is made up of 30 randomly selected questions and covers all the topics of the course. The data obtained from this questionnaire constitute the final step of the in-class role and their analysis yields what in this work has been called the long-term results of the online activities. All these students' interactions with the online activities are tracked by Moodle and allow us to extract value from the data through data analytic, as shown in the next section.

Table 1. Centroids of each Cluster (mean,std.dev)

Parameter	Cluster0	Cluster 1	Cluster2
T	(0.3405, 0.727	(3.1199, 3.634)	(8.9489, 6.754)
Re (T max)	(0.0012,0,003)	(0.7083,1.501)	(3.1395,3.503)
R	(0.4373,1.022)	(1.3081,1.935)	(3.286, 3.431
C	(0.0021,0.004)	(0.0159,0.020)	(0.0749, 0.101)

a. Data analysis results

The results shown in this section have been obtained using Weka [27]. First, we have applied the expectation-maximization (EM) clustering algorithm. This algorithm sorts the students into three groups that we will call Cluster0, Cluster1 and Cluster2. Each cluster includes a different number of students but with similar behavior. Table 1 shows this information and both the centroids and standard deviation of the behavioral parameters are described in Section 4 for each cluster. Note that the centroid value does not have to represent the behavior of any student but describes the most typical case within the cluster. Students in Cluster0 have the lowest iteration values with online activity. In Cluster0, the average number of times the activity is performed in the training period, T, is less than 1. Moreover, in this cluster the short time spent performing the activity (C parameter) is striking. On the other hand, it can be shown that the students composing Cluster2 are quite active in the training period. In Cluster2, students make use of the training activity a high number of times (T=8) and also dedicate more time to the training process (high values in C and R). In Cluster1, students also show active participation in online activities, T=3, although with lower values than Cluster2.

First, we will analyze if students' behavior associated with Cluster0, Cluster1 and Cluster2 affects students' cognitive outcomes. Figures 3-a and 3-b show the distribution of students in each cluster according to the cognitive values RS and RF, respectively. Figure 3-a shows the results of the short-term learning process, and it can be observed that students in Cluster2 obtain the best results. More than 85% obtain an excellent result. Furthermore, only students from Cluster0 fail the activity. The higher the participation in the training, the better the academic results. These results highlight the effectiveness of the online activities in the short-term.

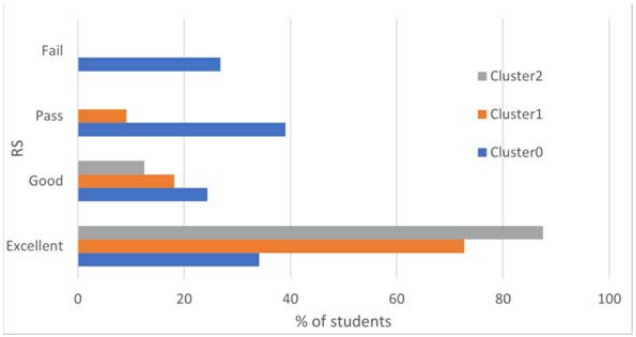

Figure 3-a. Cognitive distribution of students according to the EM algorithm (RS distribution)

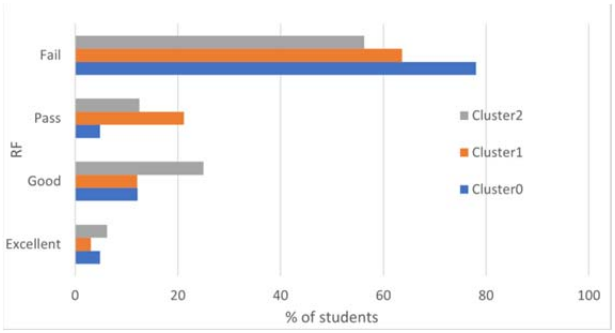

Figure 3-b. Cognitive distribution of students according to the EM algorithm (RF distribution)

In Figure 3-b the FS distribution, deeply linked to the long-term teaching-learning process, is shown. In this case, the high percentage of students who do not pass the test stands out. So, results show that over time, learners have forgotten what they learnt in the online training phase. However, students in Cluster2 are less likely to fail. At this point it should be mentioned that the complexity of the final test is greater than those taken in the short-term learning process, as it covers the content of all the topics of the subject. Although this aspect may affect the results obtained, the impact of online learning activities on long-term academic results is limited.

To continue the search of valuable information, data are analyzed applying decision trees. The results of this analysis identify which behavioral parameter has the greatest impact on academic results. Thus, the identification of inappropriate behavior will allow us to take corrective educational actions and improve the quality of teaching. This analysis aims to identify the behavioral parameters that have the greatest impact on academic outcomes to be able to apply corrective measures in future implementations and improve the quality of online activities. This process comprises two parts: analyzing the short-term teaching-learning process which takes into account RS and the long-term learning process directly related to the FS attribute. Four datasets are evaluated: the three datasets obtained from EM algorithm called as Cluster0, Cluster1 and Cluster2 and also the dataset that includes all the students, hereinafter called as ClusterAll.

1) Short-term teaching-learning process DM analysis: First, we execute the J48 classification algorithm using the behavioral attributes and consider RS as the objective attribute with four nominal values (Fail, Pass, Good and Excellent)

Table 2 summarizes the main characteristics (precision, number of nodes and leaves) when the J48 classification algorithm is applied, and RS is the objective attribute to predict. The EM clustering algorithm has an efficient effect on the classification algorithm reducing the complexity of the decision tree.

Table 2. Output performance of the J48 algorithm for RS as classification attribute

Parameter	ClusterAll	Cluster0	Cluster 1	Cluster2
Precision	83.33	70.73	81.81	87.5
Nodes	10	4	5	1
Leaves	11	5	7	1

Figure 4 shows the decision trees obtained for ClusterAll and Cluster0. Results reflect that R, T and their standard deviation are the key parameters. A high value of R reports an Excellent outcome (see Figure 4-a). An important finding is that T is a key attribute. All the Fail outcomes are associated to students who were not involved in the training step. Another finding is that Re does not appear in the classification algorithm for any dataset. Moreover, C is not a decisive attribute and, beyond any logical reasoning the classification algorithm shows that students with less cost value get better outcomes.

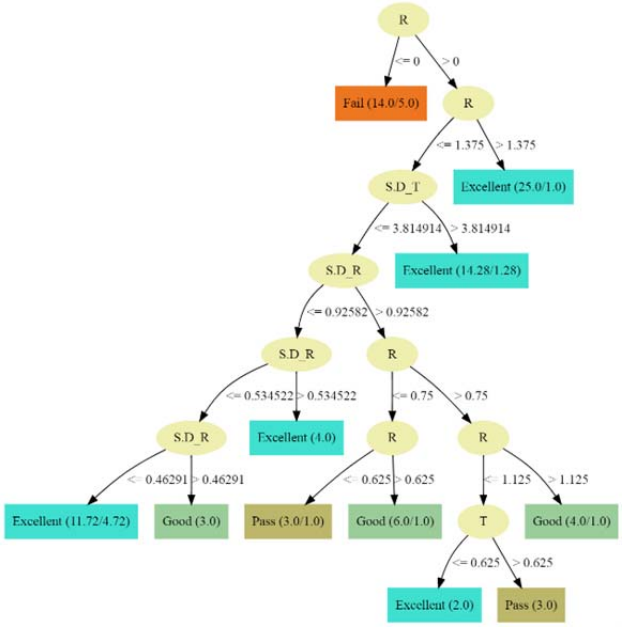

Figure 4-a. Classification of students for the short-term learning process using the J48 algorithm (ClusterAll)

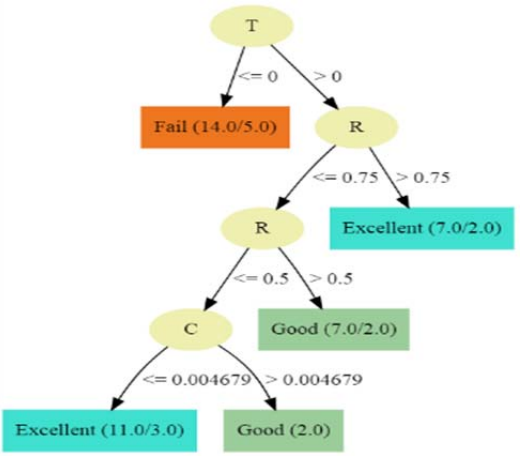

Figure 4-b. Classification of students for the short-term learning process using the J48 algorithm (Cluster0)

2) Long-term teaching-learning process DM analysis: In this case, the FS parameter will be used to discover the long-term learning process model, as well as to figure out how to improve the process from the student's point of view.

Table 3 summarizes the results of applying the J48 algorithm to predict RF with the different datasets and Figure 5 shows the decision trees obtained.

Table 3. Output performance of the J48 algorithm for RF as classification attribute

Parameter	Cluster0	Cluster 1	Cluster2
Precision	87.8	78.78	87.5
Nodes	3	4	4
Leaves	6	5	5

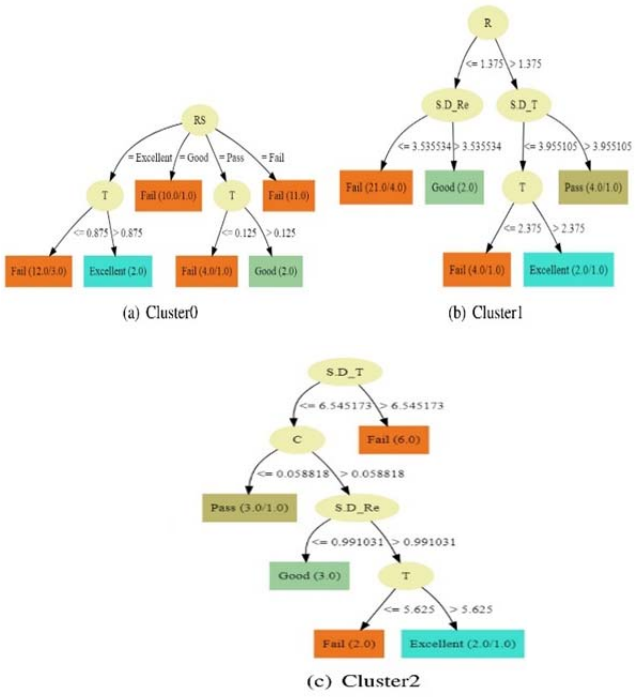

Figure 5. Classification of students for the log-term learning process using the J48 algorithm

In Cluster0, students obtain the highest scores in FS only if they have a high score in RS and participate actively in the training step (high value in T). It is remarkable that students with RS=Excellent but low participation in the training step fail the final step. Data show that most of them are students that enroll again in the course. To sum up, students of Cluster0 achieve good final scores, FS, if they are involved in the training step and attain good results in the short-term learning process. The decision tree obtained for Cluster1 students shows that R, T and the standard deviation of T and Re are the attributes used in its internal nodes to classify the target nominal values of FS (see Figure 5-b). Moreover, Cluster2 students are classified according to T, C and the standard deviation of T and Re (see Figure 5-c). Note that Cluster2 is the most active group on the online activity. This group fails with a high standard deviation of T and reaches the best score with a low standard deviation in T and high T.

To sum up, clustering techniques simplify the complexity of the models enabling better data analysis but, long-term outcomes presents different behavioral patterns being not only related to academic short-term outcomes.

7. Conclusions and Future Work

Tracking and analyzing students' activity is critical in online activities. In this work, behavioral and cognitive parameters of online activities have been defined. And techniques based on DM have uncovered significant patterns in students' behavior. These patterns are key to recognizing trends among students' work and improving their learning outcomes. The data analysis shows that integrating online training activities into the learning process improves students' results, especially in the short term. Moreover, the parameter that most influences the cognitive outcomes is the number of times a training activity is completed. It is important to highlight that the methodology has been clearly exposed, and open-source tools has been used to simplify the comparison and replication of similar experiences for future improvements. The datasets extracted from our online activities using Moodle quizzes are available to the scientific community.

We believe that online training activities could be improved by integrating a recommendation mechanism generated from the behavioral metrics of each student and the results obtained in this work. Therefore, we would like to extend the online activities with a real-time notification system that guides students in the in-class role.

References

[1]. Mishra, L., Gupta, T., & Shree, A. (2020). Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *International Journal of Educational Research Open*, *1*, 100012.

[2]. Kahoot (2022). Kahoot, learning games, make learning awesome! Retrieved from: https://kahoot.com/. [accessed: 23 December 2022].

[3]. Socrative (2022). Meet Socrative: Your classroom app for fun, effective engagement and on-the-fly assessments. Retrieved from: https://www.socrative.com/ [accessed: 11 January 2022].

[4]. Moodle (2022). Moodle: Community driven, globally supported. Retrieved from: https://moodle.org/, [accessed: 11 January 2022].

[5]. Martin Nunez, J. L., Tovar Caro, E., & Hilera Gonzalez, J. R. (2017). From Higher Education to Open Education: Challenges in the Transformation of an Online Traditional Course. *IEEE Transactions on Education*, *60*(2), 134-142.

[6]. Kangas, M., Koskinen, A., & Krokfors, L. (2017). A qualitative literature review of educational games in the classroom: the teacher's pedagogical activities. *Teachers and Teaching*, *23*(4), 451-470.

[7]. Martínez-Ortiz, I., Pérez-Colado, I., Rotaru, D. C., Freire, M., & Fernández-Manjón, B. (2019, April). From heterogeneous activities to unified analytics dashboards. In *2019 IEEE global engineering education conference (EDUCON)* (pp. 1108-1113). IEEE.

[8]. Yassine, S., Kadry, S., & Sicilia, M. A. (2016, April). A framework for learning analytics in moodle for assessing course outcomes. In *2016 ieee global engineering education conference (educon)* (pp. 261-266). IEEE.

[9]. Brooks, D. C., & Pomerantz, J. (2017). ECAR Study of Undergraduate Students and Information Technology, 2017. *EDUCAUSE*.

[10]. Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, *68*(4), 1961-1990.

[11]. Zytkow, J. M., & Klösgen, W. (2002). Multidisciplinary contributions to knowledge discovery. In *Handbook of data mining and knowledge discovery* (pp. 22-32).

[12]. Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683-695.

[13]. Li, L. Y., & Tsai, C. C. (2017). Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance. *Computers & Education*, *114*, 286-297.

[14]. Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301-309.

[15]. 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.

[16]. You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, *29*, 23-30.

[17]. Çebi, A., & Güyer, T. (2020). Students' interaction patterns in different online learning activities and their relationship with motivation, self-regulated learning strategy and learning performance. *Education and Information Technologies*, *25*(5), 3975-3993.

[18]. Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991-16005.

[19]. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199).

[20]. Antonaci, A., Klemke, R., & Specht, M. (2019, September). The effects of gamification in online learning environments: A systematic literature review. In *Informatics* (Vol. 6, No. 3, p. 32). Multidisciplinary Digital Publishing Institute.

[21]. Yang, Y., Hooshyar, D., Pedaste, M., Wang, M., Huang, Y. M., & Lim, H. (2020). Prediction of students' procrastination behaviour through their submission behavioural pattern in online learning. *Journal of Ambient Intelligence and Humanized Computing*, 1-18.

[22]. Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. *International Educational Data Mining Society*.

[23]. Martinez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39-55.

[24]. Recker, M., & Lee, J. E. (2016). Analyzing learner and instructor interactions within learning management systems: Approaches and examples. *Learning, Design, and Technology*, 1-23.

[25]. Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. *Data mining and learning analytics: Applications in educational research*, *1*.

[26]. Univ. Castilla-La Mancha, (2022). Campus virtual University of Castilla-la Mancha. Retrieved from: https://www.uclm.es/areas/areatic/servicios/docencia/campusvirtual, [accessed: 15 January 2022].

[27]. WEKA (2022). Weka: The workbench for machine learning. Retrieved from: https://www.cs.waikato.ac.nz/ml/weka/, [accessed: 16 January 2022].