Research in Engineering Education Symposium &
Australasian Association for Engineering Education Conference
5 – 8 December, 2021 - Perth, WA

Process Mining Model to visualize and analyze the Learning Process

Maria Moreno[a], Ernesto Exposito[a], and Mamadou Gueye[a]
[a]*Université de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA, Anglet, France*
Corresponding Author's Email: m.moreno-exposito@etud.univ-pau.fr

CONTEXT

In online learning environments, the teacher provides students with a learning path to follow in order to acquire the expected competencies and skills. However, students' profiles are different as they can learn according to different learning paces or media content. Therefore, the actual learning path followed by each learner may vary from the initial path provided in the learning management system (LMS). This paper proposes an analysis of the learning paths followed by the students in order to identify and promote the most adapted learning processes in order to improve competencies and skills acquisition.

PURPOSE OR GOAL

The learning traces left by students in their learning environment could be exploited in order to better understand and guide learning processes. Unfortunately, with large-scale education, the analysis of different learning paths can be a complex task to be manually carried out by teachers. For this reason, our objective is to propose an approach to model, visualize, analyze and recommend the most efficient learning process in order to improve students' education experience and results.

APPROACH OR METHODOLOGY/METHODS

The approach adopted is based on the learning traces left by the students following their interactions with the Learning Management System (LMS). After collecting, processing, and storing these learning traces, Process Mining technologies are used to analyze the data through an exploration of the learning process, as well as the students' learning paths.

ACTUAL OR ANTICIPATED OUTCOMES

The first results obtained have made it possible to visualize the learning process, as well as the learning paths followed by each learner. They also provide analysis indicators for understanding and optimizing the learning process and the students' paths in digital learning environments. These results allow the stakeholders (training managers, teachers, and students) to improve the way they teach and learn.

CONCLUSIONS/RECOMMENDATIONS/SUMMARY

This approach made it possible to comprehensively understand the learning processes and the learning paths of each learner, to visualize their differences, as well as their advantages and disadvantages. The analysis of the learning processes promoted a correlation study between the behavior of the learner (i.e. the number of connections between the sections of a course followed) during the learning process and their mark obtained on the final exam. The correlation coefficient of the evaluated courses was of the order of 0.49 and 0.53 respectively. Moreover, in order to improve the predictive model, it's necessary to implement advanced analysis: diagnosis, predictive and prescriptive based on the descriptive elements (Process visualization). This allows teachers to have an integrated tool for analyzing learning traces through a monitoring, diagnostic, alert, and early intervention system in order to better promote the success of students.

KEYWORDS

Learning Analytics, Learning Process, Process Mining, Learning Paths, Online Education

Introduction

Nowadays, online learning is largely present in educational institutions around the world. In online learning environments, teachers provide students with a learning path to follow in order to acquire the competencies and skills related to the courses. However, students present different profiles and they learn according to different learning paces and can interact differently face to heterogeneous media content. Therefore, the final learning path followed by each learner may vary from the initial path provided in the learning management system (LMS). The learning traces left by students in their learning environment could be exploited by teachers for an improvement of the learning, but unfortunately, with large-scale education, the analysis of different learning paths can be a complex task to be carried out manually by teachers. In the Learning Analytics domain, several solutions applying different methods to improve learning processes have been developed to visualize data about students and their performance or to generate recommendations using prediction models to improve the decision-making process. Other solutions have also been proposed in order to manage adaptive feedback and support collaborative argumentation in F2F (*face to face*) contexts. In contrast, studies oriented to better understanding how exactly the learning process occurs have not been extensively carried out. Nevertheless, an integrated approach aimed at analyzing students' behavior patterns, could offer interesting benefits by combining both process mining and learning analytics techniques. In this paper, a Process Mining Model is proposed to visualize and analyze the learning process, as well as the learning paths followed by the students in order to globally improve students' results. This will consist of the generation of a step-by-step modeling process to guide the implementation of learning analytics techniques based on process mining: starting with the collection of data (definition of sources), the storage of data (Learning Record Store), treatment of data (necessary transformations to be able to implement the process mining), analysis of data (application of process mining) and finally, data visualization (the respective analyses "process discovery"). The paper is organized as follows: Section II describes the Background. Section III summarizes Related Works. Section IV presents the proposal for the process mining model. Finally, section V concludes the paper.

Background

In the context of learning environments, every interaction made by the stakeholders (students, educators, institutions), leaves important traces of information that can be recorded, obtaining large sets of educational data or *Big Data*. The term *Big Data* is referred to as datasets whose size is beyond the ability of a typical database software tool to capture, store, manage and analyze (Manyika et al., 2011). The capability to extract value from such datasets is the work of Learning Analytics, defined as the application of analytics to enhance or improve student success, as well as the use of data, statistical analysis, and explanatory and predictive models to gain insight and act on complex issues (Arroway et al., 2016).

Learning Analytics is still in its infancy; however, its short life has produced numerous conceptualizations (Munguia et al., 2020). It has also helped with the development and implementation of tools that allow institutions to monitor and understand their students and the barriers to student learning. To provide a better understanding for educators of how their content is being used and how effective it is in favor of enabling its continual enhancement. Also enabling students to take control of their learning to give them better information on how they are progressing and what they need to do to meet their educational goals (Leitner et al., 2017).

Students within an LMS follow a learning path, defined as the implementation of a curriculum design; it consists of a set of learning activities that help users achieve particular learning goals (Nabizadeh et al., 2020). However, all the students due to their different profiles, backgrounds, and levels of knowledge, have particular behaviors, and accordingly, the resulting learning path can differ across students. All the data generated through these interactions can be stored implementing xAPI (*Experience API)* which is a technical

specification that aims to facilitate the documentation and communication of learning experiences (Advanced Distributed Learning (ADL) Initiative, 2013)

All the stored data can be analyzed, bringing benefits not only in terms of learning research but also in terms of didactics and actual teaching practice (Juhaňák et al., 2019). To accomplish this it is possible to apply process mining techniques, defined as the creation of a consistent and explicit process model given an event log and the use of tools to diagnose issues observing dynamic behavior (van der Aalst, 2016). Process mining is an emerging discipline providing comprehensive sets of tools to provide fact-based insights and to support process improvements, this new discipline builds on process model-driven approaches and data mining. The goal of process mining is to use event data to extract process-related information, to automatically discover a process model by observing events recorded by some enterprise system. This field has had important developments and their main goals are focused on offering ways to automate some tasks integrated into a human task and to control the information flow. (Aalst, 2011).

Related Works

One of the approaches was developed by Gutiérrez et al., 2020, which includes the implementation of LADA, a Learning Analytics dashboard to help advisers in the decision-making process, thanks to the delivery of predictions in terms of percentage of students academic risk based on an Adaptive Multilevel Clustering technique using data from the past such as previous academic records, and data from the present like the specific course selected by the student. The success component uses these estimations translating the percentage of risk in the likelihood of student success towards an individual course or a group of selected courses.

Additionally, (Han et al., 2020) developed a dashboard system for both students and instructors to support “collaborative argumentation”. The student dashboard delivered benefits such as monitoring current learning status, receive adaptive assessing from the teachers and support for FCA (face-to-face collaborative argumentation), and allow the possibility to ask for help from the teacher. The teachers’ dashboard benefits were related to monitoring the general performance of the class and identifying groups that needed help, which allowed teachers to improve decision-making regarding selection and preparation of the support to give to students based on their respective needs.

Furthermore, Aljohani et al., 2019 presented a “course-adapted student learning analytics framework” that had four different levels. Instructor level, data level, data analytics level, and presentation level (data visualization). The instructor level, included some configurations steps, being the first to specify the tools from the LMS to apply into the course; the second was to specify the data of interest, based on the chosen LMS tools; and finally in order to communicate with students could be implemented emailing process or posting in the announcement area of the LMS. As for the data level, this aimed to extract and retrieve the data from LMS that relates to the LMS tools being used by students for the course, to be later analyzed in the data analytics level employing several techniques. Regarding data visualization, the AMBA tool was developed which is a web-based application; teachers and students had access to the tool; as for the teachers they were able to configure the tool, and as for the students, the tool provided three types of feedback in the order of statistical, textual and visual.

Additionally, Juhaňák et al., 2019 carried out a study, centered on student behavior in LMS (in this case, the widely used open-source system Moodle), specifically on student interactions while engaging in specific quiz-based learning activities. The analysis of student interactions uses process mining methods, allowing for mapping and modeling the process of completing quizzes by students.

The following Table 1 summarizes the approaches of the sources consulted, in the different steps identified to implement Learning Analytics.

Table 1: Learning Analytics Implementation Comparison

Features	Sources				
	(Gutiérrez et al., 2020).	(Han et al., 2020)	(Aljohani et al., 2019)	(Juhaňák et al., 2019)	Proposal
Data Collection	YES	YES	YES	YES	YES
Data Storage	YES	YES	YES	YES	YES
Data Treatment	NO	YES	YES	YES	YES
Data Analysis	Multilevel Clustering	Statistical Analysis	Unclear	Process Mining	Process Mining
Data Visualization	YES	YES	YES	NO	YES

Process Mining Model

Model Overview

The main goal of this model is to generate a generic approach. A step-by-step guide that helps institutions acknowledge all the concepts related to the implementation of a Process Mining model to visualize and analyze the learning process that takes place on an LMS (see Figure 1).

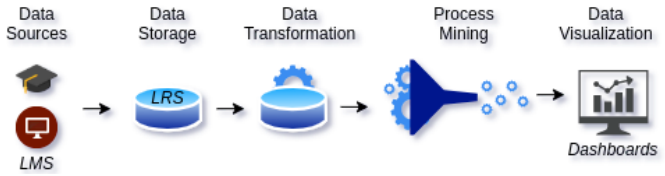


Figure 1: Process Mining Model to visualize and analyze the learning process.

Data Sources

In order to visualize the learning process, it is necessary to extract all the relevant information referring to how these processes occur, and the places where such activities happen are systems like OLE *Online Learning Environment* or LMS *Learning Management Systems*. In those systems each click, view, answer, success, error, time consumed, resource downloaded, viewed, listened or score obtained, has an important meaning, and all of those interactions build some digital footprint, also called Learning Records which are the one's vitals to collect.

Data Collection

To collect, extract and store the interaction data, it is necessary to follow a standard approach like the one proposed byTinCanAPI or xAPI (Experience API) which is a technical specification that aims to facilitate the documentation and communication of learning experiences. In general, when an activity needs to be recorded, the application sends secure statements in the form of *noun, verb, object* or *I did this* to a Learning Record Store (LRS) which is a server that is responsible for receiving, storing, and providing access to Learning Records. In general (Advanced Distributed Learning (ADL) Initiative, 2013) explains that its necessary to collect the following:

- **Actor**: is an individual or group representation tracked using Statements performing an Action within an Activity. Is the “I” in “I did this”.
- **Verb**: Is the action being done by the Actor within the Activity within a Statement. A Verb represents the “did” in “I did this”.

- **Activity**: a type of Object making up the “this” in “I did this”; it is something with which an Actor interacted. It can be a unit of instruction, experience, or performance that is to be tracked in a meaningful combination with a Verb.

Data Treatment

The process mining analysis starts with an ‘Event Log’ and inside this structure, a process is described as follows, a process consists of *cases,* a case consists of *events* such that each event relates to precisely one case and each *sequence* of activities executed for a case is a *trace*. Each line in the event log presents one event. Events within a case are *ordered.* Events can have *attributes (e.g.,* activity, time, cost, resource, etc) (Aalst, 2011). The event log structure can be summarized as follows:

- **Case ID**: indicates at which case or instance belongs an event or activity.
- **Activity**: Action captured by the event.
- **Timestamp**: Indicate the time when the event took place.

Concerning the proposed model, the first field to build the “Case ID”, should include information that allows a unique identification of the actor (student), like the student ID. The second field is the activity field, this can be obtained using the “activity” of the previously stored learning experiences. Regarding the timestamp field, the easiest approach is to duplicate the timestamp information of each learning record into the respective event log. Table 2 displays an example of a construction of an event log from trace data of a course called “Advanced Databases” followed by twenty-one students (some interactions of two students are displayed). Additionally, it is possible to record other information considered important, which subsequently will function as filters in the analysis phase, the “verb” of the learning experiences previously-stored the could be used. Another interesting filter could be the grades associated with the student, however, this isn't necessary to add it in the events logs, on the contrary, this information could be stored in a separate file with a simple structure that contains the “Case ID”, and the respective grade; Table 3 describes an example of this (20.0 scale). At the analysis phase, it will be possible to filter by grades, adding the grades table to the analysis and connecting the events logs to the grades table using the “Case ID” as a foreign key.

Table 2: Event log construction example from trace data

case_id	activity	timestamp	status
55138	Data Integration Introduction	12/01/21 08:23	viewed
55138	Data Integration Introduction	12/01/21 08:23	completed
67108	Data Integration Introduction	12/01/21 08:32	viewed
...	...	...	...
67108	Lab 1 - Data Integration	12/01/21 09:07	viewed

Table 3: Grades file

case_id	final_grade
67256	17,539
67108	15,091
55138	13,746
...	...
67239	6,672

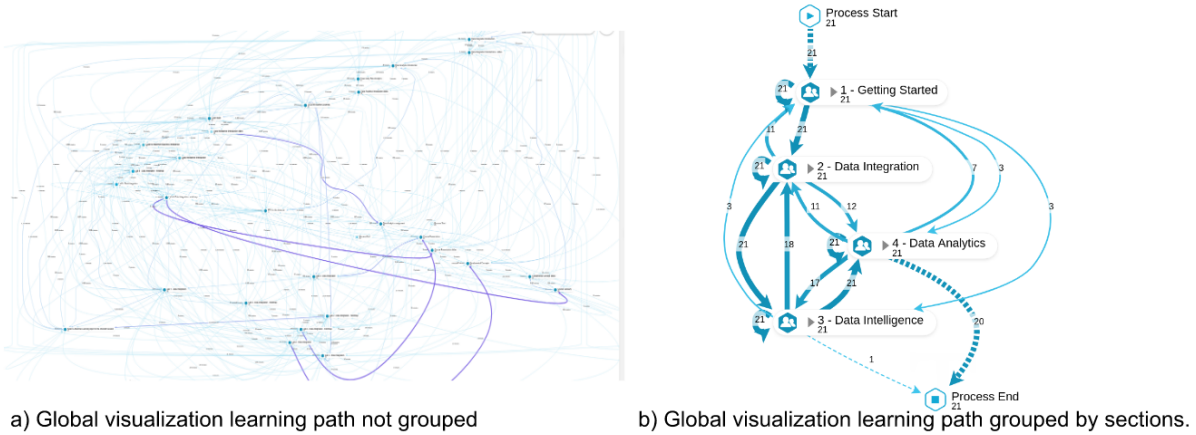
Data Analysis

Once the data have been transformed to the necessary structure ‘event log’, one of the most suitable algorithms to implement in this step is the *“discovery”* type. This technique takes an event log and produces a model without using any a-priori information (Aalst, 2011). It is important to point out that the actual implementation of this algorithm is not specified by this model, thus it can vary from one implementation to another. However, it is strongly recommended to use an online existing solution that already implements process mining, for example, the Celonis platform (*Process Mining and Execution Management Software*, n.d.) to speed up the process. However, regardless of the final implementation once applied to the discovery algorithm, the discovery model should be “representative” for the behavior seen in the event log (Aalst, 2011).

To summarize some benefits of applying process mining to educational event logs (from trace data), hereafter these are going to be described focused on the three main stakeholders of a learning process (educators, training managers, and students).

Benefits for educators

- **Visualize globally the learning process:** with this kind of analysis, it is possible to visualize globally all the interactions made in the learning process by all students (Figure 3a). At first, this visualization is too complex to be analyzed, however, it is possible to group the activities by sections (several activities can be part of the same section or learning outcome), with this approach it is possible to globally visualize all the interactions of the students grouped as needed (Figure 3b). Case frequency is also displayed, which applied to the current context represents the number of students that follow a specific path. In (Figure 3b), it is possible to visualize that twenty-one students started the process and went to perform the activities of the first section "1 — Getting started". After that the twenty-one students went to the second section "2 — Data Integration", being in this section eleven students went back to the first section "1 — Getting started". In summary, it gets visually described all the sequence of resource usage by students, it is possible to see not only the straight link between resources but also all the turns, reuses and go-backs, performed in the way.



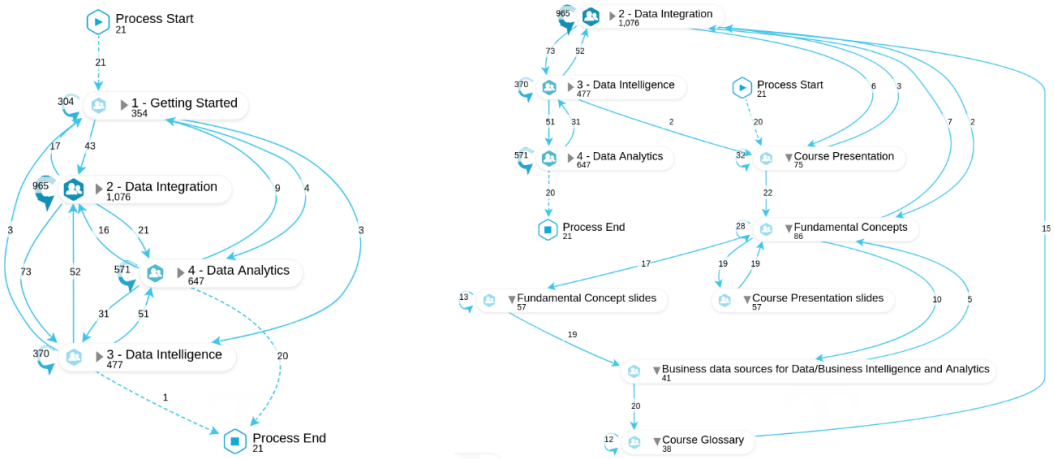
a) Global visualization learning path not grouped

b) Global visualization learning path grouped by sections.

Figure 3: Learning process. Case Frequency - Process explorer Celonis platform.

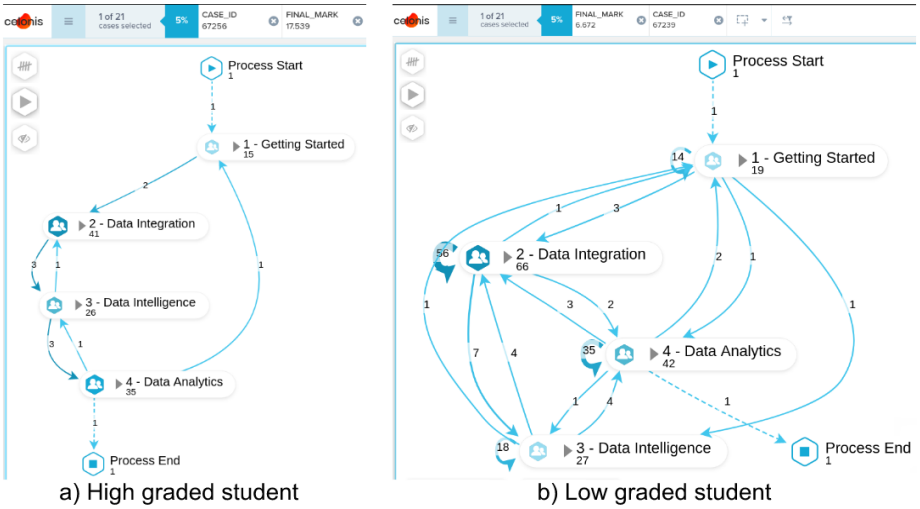
- **Visualize frequency of resource usage**: the activity frequency can also be displayed (Figure 4a), which translates to the number of times an activity was performed. This can be helpful at the time of evaluating the resources usage and value. Several assumptions can be performed based on the number of times a resource is being used; depending on the context this could lead to improvement of the quality of resources if they are considered as difficult for students. As previously stated the visualization can be generally grouped by sections (Figure 4a), or detailed sections expanded (Figure 4b).

- **Visualize the learning process of an individual student or group of students:** it is possible to filter the cases (students), to visualize the individual performance or even subgroup performance. The filter can also be performed by grades to help identify the paths taken by students that led to higher or lower grades. Also, could help to identify patterns in behavior to be able in the future to predict if a student could be at risk of obtaining a bad grade or even dropping out. Another important benefit is the possibility to visualize students having problems with specific subjects. For example, Figure 5a and Figure 5b, represent the learning path taken by two different students who obtained a grade of 17.539 over 20 and 6.672 over 20 respectively. The big disparity between these two students' grades can be visualized also in the path taken. In Figure 5a it is possible to see a more organized path, with fewer connections between sub-learning paths of the course. On the contrary, Figure 5b, represents a learning path more disorganized, with several connections between sub-learning paths of the course.



a) Global visualization - Learning path grouped by sections. b) Detail visualization per sections of a the learning path

Figure 4: Learning process. Activity frequency - Process explorer Celonis platform.



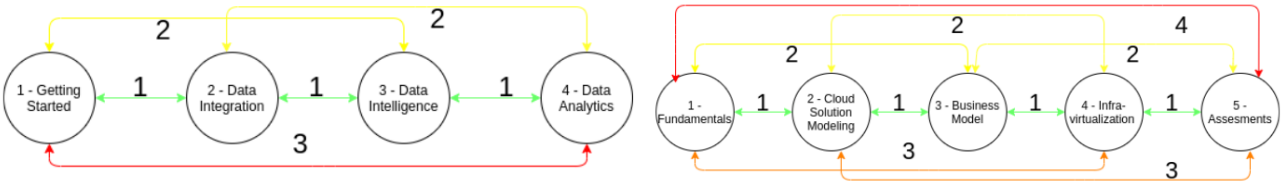
a) High graded student b) Low graded student

Figure 5: Individual Learning process grouped by sections. Case frequency - Process explorer Celonis platform.

Benefits for training managers

- **Predictions of students' performance:** In order to predict student's performance, it was necessary to apply a more detailed study of the individual cases. In general, two courses were analyzed, called "Advanced databases" and "Cloud Computing II" respectively. For each of the courses a generic path was constructed, in Figure 6a we see an example of this generic path for the course "Advanced databases". With this, it is possible to visualize the weights between sections of a learning path. We hypothesize that the connections between sections directly related should weigh 'one', while the connections between sections not directly related should weigh the sum of the previous weights. For example the weight of the connections of the sections "1 - Getting Started" and "2 - Data Integration" has a value of 'one', however, the connection between the sections "1 - Getting Started" and "3 - Data Intelligence" or vice-versa has a value of 'two' which is the sum of the two previous steps. This is proposed with the idea of carrying out a linear regression model taking as an independent variable the sum of weight per connections of students' individual paths and as the dependent variable the final grade. As a result, it is possible to predict the grades of the students based on their interactions on the learning management system, analyzing their learning paths and identifying patterns. This study was performed with two courses and Table 4 represents the results obtained and the

corresponding correlation coefficient, in both cases the coefficient represents an average value, in consequence further testing with more courses data need to be analyzed in order to improve the predictive model, based on each particular course.



a) Generic path. Course Advanced Databases

b) Generic path. Course Cloud Computing II

Figure 6: Generic path with weights between sections.

Table 4: Linear regression on the number of connections of students' individuals' paths

Course	Final grades predictive model	Correlation coefficient
Advanced Databases	-0.13 * connections + 15.84	0.4951
Cloud Computing II	-0.08 * connections + 19.99	0.5373

Benefits for students

- **Learning path recommendations for a course:** recommendations for the learning path are related to the one that obtained the higher grades. Based on the previous academic course, recommendations can be made to students following the same course. In Figure 7 it is possible to visualize a learning path of students that obtained grades between 18 – 20 on a scale of 20.

Figure 7: Path with high grades. Case frequency - Process explorer Celonis platform.

Limitations

This research was performed during the academic year of 2020-2021 and the data available for the courses analyzed was limited to only this period. In order to improve the results of the recommendations and predictions, it's necessary to analyze more data from the same courses but in another academic period.

Conclusions and Future work

This paper proposed the design of a Process Mining model to visualize and analyze the Learning Processes of students. The analysis is performed by collecting learning experiences from students in regular LMSs, store them on a common LRS using the xAPI standard and apply some transformations to build event logs. Those event logs are then analyzed using process mining and their results are exposed visually recreating the learning processes in their given contexts. This approach allows the possibility to explore all the different ways that students take to learn similar subjects (different learning paths); be able to visualize globally, individually, or by subgroups the learning process of each one of them, and

remark their differences, advantages, or disadvantages. The frequency of resource usage can also be visualized and each teacher can draw their own conclusions that help them evaluate the quality of their resources. Consequently, teachers and educational institutions can have access to the visualization of all this information allowing them to self-reflect on their practices and have an overview of the current situation that could help them to make the best decisions that would help improve as much as possible the learning environment processes. The analysis of the learning processes promoted a correlation study between the behavior of the learner (number of connections between the sections of a course followed) during the learning process and their mark obtained on the final exam. The correlation coefficient of the two courses studied was of the order of 0.49 and 0.53 respectively, representing an average value. In consequence, to improve the predictive model, it's necessary to implement advanced analysis: diagnosis, predictive and prescriptive based on the descriptive elements (Process visualization). This allows teachers to have a complete tool for analyzing learning traces through a monitoring, diagnostic, alert, and early intervention system to better promote the success of students.

References

Aalst, W. van der. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag. https://doi.org/10.1007/978-3-642-19345-3

Advanced Distributed Learning (ADL) Initiative. (2013). *Experience API, Specification document*. GitHub. https://github.com/adlnet/xAPI-Spec

Aljohani, N. R., Daud, A., Abbasi, R. A., Alowibdi, J. S., Basheri, M., & Aslam, M. A. (2019). An integrated framework for course adapted student learning analytics dashboard. *Computers in Human Behavior*, *92*, 679–690. https://doi.org/10.1016/j.chb.2018.03.035

Arroway, P., Morgan, G., O'Keefe, M., & Yanosky, R. (2016). *Learning Analytics in Higher Education*. 44.

Gutiérrez, F., Seipp, K., Ochoa, X., Chiluiza, K., De Laet, T., & Verbert, K. (2020). LADA: A learning analytics dashboard for academic advising. *Computers in Human Behavior*, *107*, 105826. https://doi.org/10.1016/j.chb.2018.12.004

Han, J., Kim, K. H., Rhee, W., & Cho, Y. H. (2020). Learning Analytics Dashboards for Adaptive Support in Face-to-Face Collaborative Argumentation. *Computers & Education*, 104041. https://doi.org/10.1016/j.compedu.2020.104041

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior*, *92*, 496–506. https://doi.org/10.1016/j.chb.2017.12.015

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. In A. Peña-Ayala (Ed.), *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning* (pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-52977-6_1

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). *Big Data: The Next Frontier for Innovation, Comptetition, and Productivity*.

Munguia, P., Brennan, A., Taylor, S., & Lee, D. (2020). A learning analytics journey: Bridging the gap between technology services and the academic need. *The Internet and Higher Education*, *46*, 100744. https://doi.org/10.1016/j.iheduc.2020.100744

Nabizadeh, A. H., Leal, J. P., Rafsanjani, H. N., & Shah, R. R. (2020). Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications*, *159*, 113596. https://doi.org/10.1016/j.eswa.2020.113596

Process Mining and Execution Management Software. (n.d.). Celonis. Retrieved July 15, 2021, from https://www.celonis.com/

van der Aalst, W. (2016). Process Mining: The Missing Link. In W. van der Aalst (Ed.), *Process Mining: Data Science in Action* (pp. 25–52). Springer. https://doi.org/10.1007/978-3-662-49851-4_2