

## **An analysis of students' behaviour in a Learning Management System through Process Mining**

Joana Carmo Leal

Dissertation presented as partial requirement for obtaining  
the Master's degree in Information Management

**NOVA Information Management School**  
**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa

# **AN ANALYSIS OF STUDENTS' BEHAVIOUR IN A LEARNING MANAGEMENT SYSTEM THROUGH PROCESS MINING**

by

Joana Carmo Leal

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information Systems and Technologies Management

**Advisor:** Dr. Vitor Santos

**Advisor:** Dr. Luka Tomat

February 2022

## ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Prof. Dr. Vitor Santos and Prof. Dr. Luka Tomat for their guidance, patience and all the knowledge and expertise they shared with me. This path would have been more difficult without them and their insightful recommendations and advice.

I wish to show my appreciation to Vanda França and Isabel Aguiar, from the Informatics Department of NOVA IMS, for their collaboration and availability in providing the necessary information for the development of this dissertation.

A special thank you to my family and friends, particularly to my mother, father, brother, sister-in-law and aunt. Words are not enough to express my gratitude for their unconditional support.

A last thank you to my nephew, who unknowingly gave me strength throughout this year to continue.

“One child, one teacher, one book, one pen can change the world.”

Malala Yousafzai

## **ABSTRACT**

The exponential growth and transformation of the Internet and information technology in recent years led to the development of several analytical tools. As is the case with process mining, it emerged to fulfill the need to extract and analyze information from event logs by representing it in the form of process models. Process mining is an acclaimed tool and proved crucial in several areas, from healthcare to manufacturing and finance. Nevertheless, and despite the crucial role of digital systems in supporting learning activities and generating large amounts of data about learning processes, limited research focused on process mining applied to the educational context.

Therefore, the aim of this dissertation is to apply a process-oriented approach and demonstrate the applicability of process mining techniques to explore and analyze students' behavior and interaction patterns, based on data collected from Moodle, the widely used Learning Management System. We cover definitions of process mining, education, and a detailed search of the existing literature on educational process mining during this work. Furthermore, the paper analyzes and discusses the findings of the study that combines process mining techniques, specifically process discovery implanted in the Disco tool, with cluster analysis.

Through the application of these two techniques, it was possible to recognize the relationship between the students' behavior registered in the process models and the success of the students in the course, along with the general and specific information about the students' learning paths. Besides, we obtained findings that allow us to predict the group of students at risk of failing. Finally, with the analysis of these results, we were able to provide improvement proposals and recommendations to enhance the learning experience.

## **KEYWORDS**

Process Mining; Education; Moodle; Expectation-Maximization algorithm; Learning behavior; Disco

# INDEX

|                                             |           |
|---------------------------------------------|-----------|
| <b>1. INTRODUCTION .....</b>                | <b>2</b>  |
| 1.1. BACKGROUND .....                       | 2         |
| 1.2. MOTIVATION .....                       | 3         |
| 1.3. RESEARCH GOAL .....                    | 3         |
| <b>2. LITERATURE REVIEW .....</b>           | <b>5</b>  |
| 2.1. PROCESS MINING .....                   | 5         |
| 2.1.1. <i>Concepts</i> .....                | 5         |
| 2.1.2. <i>Process Discovery</i> .....       | 7         |
| 2.1.3. <i>Process Conformance</i> .....     | 7         |
| 2.1.4. <i>Process Enhancement</i> .....     | 8         |
| 2.1.5. <i>Tools &amp; Algorithms</i> .....  | 9         |
| 2.1.6. <i>Guiding Principles</i> .....      | 10        |
| 2.1.7. <i>Challenges</i> .....              | 10        |
| 2.2. EDUCATION .....                        | 12        |
| 2.2.1. <i>Concepts</i> .....                | 12        |
| 2.2.2. <i>Educational Processes</i> .....   | 14        |
| 2.3. PROCESS MINING IN EDUCATION .....      | 15        |
| 2.3.1. <i>Search Strategy</i> .....         | 16        |
| 2.3.2. <i>Analysis and Discussion</i> ..... | 21        |
| <b>3. METHODOLOGY .....</b>                 | <b>24</b> |
| 3.1. PM <sup>2</sup> METHODOLOGY .....      | 24        |
| 3.2. RESEARCH STRATEGY .....                | 25        |
| <b>4. STUDY .....</b>                       | <b>27</b> |
| 4.1. OVERVIEW .....                         | 27        |
| 4.2. DATA COLLECTION .....                  | 27        |
| 4.3. DATA PROCESSING .....                  | 28        |
| 4.3.1. <i>Data Selection</i> .....          | 28        |
| 4.3.2. <i>Econometrics</i> .....            | 28        |
| 4.4. PROCESS MINING AND ANALYSIS .....      | 32        |
| 4.5. EVALUATION .....                       | 40        |
| 4.6. DISCUSSION .....                       | 45        |
| 4.6.1. <i>Evidence and Outcomes</i> .....   | 45        |
| 4.6.2. <i>Improvement Proposals</i> .....   | 47        |
| <b>5. CONCLUSION .....</b>                  | <b>49</b> |
| 5.1. SYNTHESIS OF THE RESEARCH .....        | 49        |
| 5.2. LIMITATIONS AND FURTHER RESEARCH ..... | 50        |
| <b>REFERENCES .....</b>                     | <b>51</b> |
| <b>APPENDICES .....</b>                     | <b>58</b> |

## LIST OF FIGURES

|                                                                                                                                                                                                                     |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1 - Positioning of the three main types of process mining: discovery, conformance, and enhancement (van der Aalst, 2016) .....                                                                               | 6  |
| Figure 2 - The three basic types of process mining explained in terms of input and output: (a) discovery, (b) conformance checking, and (c) enhancement (van der Aalst, Adriansyah, de Medeiros, et al., 2012)..... | 9  |
| Figure 3 - PRISMA flow diagram.....                                                                                                                                                                                 | 16 |
| Figure 4 - Search Strings.....                                                                                                                                                                                      | 17 |
| Figure 5 - Percentage of papers published per year .....                                                                                                                                                            | 20 |
| Figure 6 - Tools used in research .....                                                                                                                                                                             | 23 |
| Figure 7 - An overview of the PM <sup>2</sup> methodology (van Eck et al., 2015) .....                                                                                                                              | 24 |
| Figure 8 - Calendar heatmaps of students' activity .....                                                                                                                                                            | 32 |
| Figure 9 - Disco's Statistics view.....                                                                                                                                                                             | 33 |
| Figure 10 - Process model of all students with 100% activities and 90% paths .....                                                                                                                                  | 34 |
| Figure 11 - Process model of all students with 100% activities and 40% paths (case frequency view) .....                                                                                                            | 35 |
| Figure 12 - Process model of all students with 100% activities and 40% paths .....                                                                                                                                  | 36 |
| Figure 13 - Process model of passing students with 100% activities and 40% paths.....                                                                                                                               | 37 |
| Figure 14 - Process model of failing students with 100% activities and 40% paths .....                                                                                                                              | 37 |
| Figure 15 - Overview badge of quiz summary viewed activity from passing students (left) and failing students (right) models .....                                                                                   | 38 |
| Figure 16 - Process model of Cluster 0 with 100% activities and 40% paths .....                                                                                                                                     | 39 |
| Figure 17 - Process model of Cluster 1 with 100% activities and 40% paths .....                                                                                                                                     | 39 |
| Figure 18 - Process model of Cluster 2 with 100% activities and 40% paths .....                                                                                                                                     | 40 |
| Figure 19 - Process model of all students, academic year 2020/2021 (case frequency view) .....                                                                                                                      | 42 |
| Figure 20 - Process model of Cluster 1, academic year 2020/2021 .....                                                                                                                                               | 43 |
| Figure 21 - Process model of Cluster 3, academic year 2020/2021 .....                                                                                                                                               | 44 |
| Figure 22 - Process model of Cluster 2, academic year 2020/2021 .....                                                                                                                                               | 44 |
| Figure 23 - Process model of repeating students, academic year 2020/2021 .....                                                                                                                                      | 45 |

## LIST OF TABLES

|                                                                                                |    |
|------------------------------------------------------------------------------------------------|----|
| Table 1 - Records Summary.....                                                                 | 17 |
| Table 2 - Description of attributes from Moodle event log.....                                 | 27 |
| Table 3 - Description of the activities considered relevant to the students' performance ..... | 29 |
| Table 4 - Description of the variables used for clustering .....                               | 30 |
| Table 5 - Statistics (means $\pm$ std. dev.) of each of the three clusters .....               | 31 |
| Table 6 - Statistics (means $\pm$ std. dev.) of each of the four clusters .....                | 41 |

## LIST OF ABBREVIATIONS AND ACRONYMS

|               |                                                                    |
|---------------|--------------------------------------------------------------------|
| <b>EDM</b>    | Educational data mining                                            |
| <b>e.g.</b>   | exempli gratia                                                     |
| <b>EM</b>     | Expectation-Maximization                                           |
| <b>EPM</b>    | Educational process mining                                         |
| <b>ISCED</b>  | International Standard Classification of Education                 |
| <b>LMS</b>    | Learning Management System                                         |
| <b>LTL</b>    | Linear Temporal Logic                                              |
| <b>MOOC</b>   | Massive Open Online Course                                         |
| <b>PM</b>     | Process Mining                                                     |
| <b>PRISMA</b> | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| <b>RQ</b>     | Research Question                                                  |
| <b>SRL</b>    | Self-Regulated Learning                                            |
| <b>UNESCO</b> | United Nations Educational, Scientific and Cultural Organization   |
| <b>VLE</b>    | Virtual Learning Environment                                       |

# 1. INTRODUCTION

## 1.1. BACKGROUND

In the past decade, data science and business intelligence have become crucial subjects to support decision making and contribute to the development of several services. The increasing use of technology combined with the possibility of collecting large amounts of data motivated the growing use of data analytics systems in diverse environments. Nowadays, it is possible to affirm that data and its analysis are becoming essential for society and organizations. Besides, people are more willing to share their data, and companies are increasingly interested in obtaining and discover substantial value and reliable information from the collected data (van der Aalst, 2016). With the maturity of information systems, new functions and techniques have been developed within the process modeling context and the data-centric analysis to meet emerging needs.

Process mining is an analytic tool that combines data mining on the one hand and business process management on the other since it integrates a data-driven strategy with a process-centric approach through a combination of event logs and process models (van der Aalst, 2016). Process mining aims to acquire knowledge about processes and analyze them based on data from events available in information systems and, consequently, use the outcome to support the discussion and rationalization about the process and contribute to decision making in the current process (van der Aalst, 2016). Moreover, process mining techniques contribute to a more detailed and rigorous compliance check and verify the validity and reliability of the information about the processes (Rovani, Maggi, de Leoni, & van der Aalst, 2015).

The growing interest in data science has contributed to making process mining an emerging topic with an even more promising future. It is expected to receive greater significance and interest for, essentially, two primary reasons. First, there is a growing accessibility of information concerning the history of the processes due to Big Data systems, considering that they facilitate the collection, storage and processing of large quantities of event logs (van der Aalst, 2016). Second, there is the constant need to improve business processes intensified by the rapidly changing environment and the search for competitive advantage (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

The main process mining techniques are (i) process model discovery, which consists of the construction of complete process models from event logs, accurately reproducing the observed behavior (Trcka & Pechenizkiy, 2009), (ii) conformance checking, focus on monitoring deviations between the modeled behavior and the observed behavior based on event logs (Cairns et al., 2015), and (iii) process model extension, which is about process model improvement by projecting the information extracted from an event log associated with the same process (Bogarín, Romero, Cerezo, & Sánchez-Santillán, 2014). Therefore, process mining tools provide a more fact-based insight and contribute to making processes particularly effective and efficient. As a result, process mining has brought new ideas and energy to static process models. At the same time, it succeeds at presenting significant volumes of data collected nowadays in a process context (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

Education is inherently linked to each person's life, whether explicitly or tacitly, from the first day to the last. Education is a highly comprehensive and widely discussed concept that results in various perceptions and theories developed over the years by several areas. In a general and summarized way,

education is the process of transmitting or gaining knowledge, experience or skills, supported by the development of reasoning and judgment (Dictionary.com, n.d.). It can be about the evolution of the individual, as well as the maximum progress of society so that together they contribute to improvement and prosperity.

Education can be distinguished in two ways: first, by its narrower sense, which corresponds to the process that takes place in certain periods, from school to college, with a fixed program, classes and topics. In this case, school characterizes formal education, and it is transmitted directly and systematically; besides, the amount of education obtained is quantified by degrees and diplomas. Second, education can be defined in its broader sense when referring to all the experiences acquired throughout life along with social and personality development (Kumar & Ahmad, 2008).

## **1.2. MOTIVATION**

Today, more than ever, there is a universal awareness of the need for education. Therefore, in addition to being a human right, the duration of compulsory education has been increasing (Dulger, 2012). Nonetheless, education is a dynamic process that requires constant adaptation to each individual, society, culture, environment and historical moment. It is, therefore, necessary to comprehend that the aims of education are variable and relative. Consequently, it is essential to emphasize the need for the educational process to be flexible and adaptable to the changing living conditions (Kumar & Ahmad, 2008).

As a result of the increasingly data-driven world, the educational data mining (EDM) discipline emerged, intending to analyze data gathered from learning and teaching activities to understand the practices better. Nevertheless, EDM was not enough to reproduce all the complexity of the educational process; thus, to fill that gap and enhance understanding, it expanded to educational process mining (EPM). Through log data collected primarily from educational domains, EPM aims to discover, analyze, and model processes through a visual representation of the educational process (Bogarín, Cerezo, & Romero, 2018).

The principal motivation of this study is to use process mining techniques to understand the educational process and, consequently, find how to improve it for a superior experience for students, teachers and institutions. However, in order to reach this desired point, it is necessary to elaborate comprehensive research on the application of process mining in education. Therefore, it is essential to analyze and study the use of process mining tools in the educational field and make a detailed comparison between the existing results in the EPM area (Cairns et al., 2015). Subsequently, apply the obtained expertise to develop a study adapted to the reality of education, customizing for its needs and users (Romero & Ventura, 2013).

## **1.3. RESEARCH GOAL**

The main goal of this paper is to study and analyze the use of process mining techniques in education with the purpose of improving the quality of the process. To achieve the main goal, during the development of this research, we will attempt to answer the following research questions:

- RQ1: How effective is process mining in identifying specific student behavior patterns in the Learning Management System Moodle?

- RQ2: How does process mining contribute to predicting which students will have the most difficulty in succeeding?
- RQ3: From the process models analysis, is it possible to withdraw valuable recommendations to students and teachers towards a more personalized educational process?

## 2. LITERATURE REVIEW

### 2.1. PROCESS MINING

#### 2.1.1. Concepts

This section will introduce the most relevant concepts related to the process mining scope, along with the presentation of remarkable researchers in the field and a few theories they explored and presented.

The main process mining concepts were primarily determined by Wil van der Aalst in his book *Process Mining: Discovery, Conformance and Enhancement of Business Processes* (2011) and subsequently in the “Process Mining Manifesto” in 2012 that intended to promote process mining and define its main notions, principles and challenges (van der Aalst, Adriansyah, de Medeiros, et al., 2012). These definitions are still predominant in the scientific literature.

According to van der Aalst (2016), process mining can be understood as the bridge between data science and process science. The author defined data science as a discipline that includes machine learning, artificial intelligence and data mining. Its main objective is to turn data into real value, which can be in the form of predictions, models and data visualizations with insights to contribute to decision making. Moreover, data science incorporates the following activities: data extraction, data preparation, data exploration, data transformation and storage. Conversely, the author described process science as a field of study that combines knowledge acquired from information technology with management knowledge. It is a model-driven approach aiming to improve operational processes without relying on the facts supported by data. Instead, it requires extensive observation and manual mapping of the process steps (Reinkemeyer, 2020a). Business process management, process automation, workflow management and operations management are some subjects under the umbrella of process science (van der Aalst, 2016).

In summary, process mining fills the gap between data mining and business process management, combining event data with process models. Data mining focuses on searching patterns in large datasets, using techniques such as association rules, decision trees and clusters; however, it disregards the comprehensive picture of the end-to-end processes. On the other side, business process management as a model-driven discipline provides a view of the processes as a whole. However, it is not based on event data, and models are generally developed by hand and, therefore, have more difficulty representing the observed behavior (van der Aalst, Adriansyah, & van Dongen, 2012).

The purpose of process mining is to obtain knowledge from event logs present in information systems. The information acquired is extracted with various data analysis techniques and algorithms in order to discover, understand, monitor and improve processes (van der Aalst, Adriansyah, de Medeiros, et al., 2012). Process mining aims at characterizing complex and straightforward process flows through a comprehensive visualization, which benefited from the growing power of data processing that permits access to large amounts of events. Process mining tools allow valuable insights by providing the possibility to drill down from the simplest and aggregated view to more complex and detailed visualizations, which contributes to a more consistent picture of all processes (Reinkemeyer, 2020).

Event logs are the necessary starting point, the input data for process mining. It represents a collection of events generated and captured by information systems in the most varied circumstances (Reinkemeyer, 2020). Each event in these domains refers to an activity that is executed by some resource related to a particular time and case; additionally, it can include extra attributes to contribute with further information (van der Aalst, Adriansyah, de Medeiros, et al., 2012). Related to event logs, there is also the concept of trace, which denotes the chronological order of a particular case's activities or events (van der Aalst, 2018). In short, event logs record the activities from a process. Every execution represents a case and develops a sequence of the activity occurrences called a trace. Such occurrences are established as events and contain various attributes (Leno, Armas-Cervantes, Dumas, Rosa, & Maggi, 2018).

More recently, Lars Reinkemeyer (2020) illustrated what it means to mine a process, which begins with the identification of event logs and traces, its extraction and consequently the visualization in a form that expresses the real process flow, i.e., not assumed processes, which contributes to a higher transparency of the activities sequences by presenting them as they happened and with a comprehensive illustration (van der Aalst, 2016; Reinkemeyer, 2020).

The “Process Mining Manifesto” (van der Aalst, Adriansyah, de Medeiros, et al., 2012) presented the following three attributes to enhance the understanding of the process mining techniques. First, process mining is not exclusively about process discovery. Process discovery is one of three fundamental techniques that compose process mining, remaining conformance checking and process enhancement, as illustrated in Figure 1 (van der Aalst, 2016) and explained in detail in the following sections. In the second place, process mining is not a subdiscipline of data mining. It can use data mining techniques; however, most of them are not process-centric, making the development of new types of algorithms a necessity. Finally, the third characteristic is that even though process mining extracts knowledge from historical data, it is not limited to offline analysis, considering it allows applying the results to current cases (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

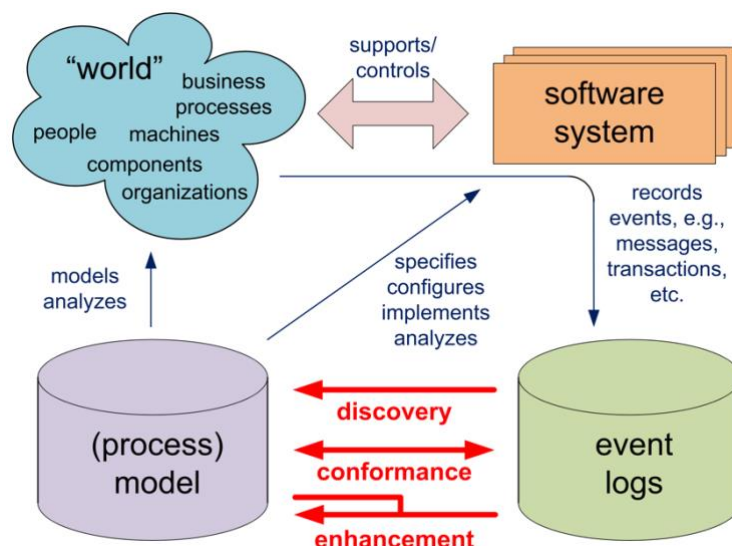


Figure 1 - Positioning of the three main types of process mining: discovery, conformance, and enhancement (van der Aalst, 2016)

As displayed in Figure 1, process mining requires a software system to store event data and, in addition, to support and control real-life processes. Furthermore, event data is extracted, filtered and loaded into the process mining software, with the intention of carrying out activities such as discovery, compliance and/or improvement (Dakic, Stefanovic, Cosic, Lolic, & Medojevic, 2018).

In recent research, Santos Garcia et al. (2019) identified the predominant exploration domains in which process mining is applied: healthcare, information and communication technology, manufacturing, education, finance and logistics. In their systematic literature review of process mining, the authors discovered some significant conclusions regarding the studies found and, consequently, how the technology is applied. They found that most research explored process discovery algorithms, followed by conformance checking, architecture and tool enhancement. Moreover, most of this research comes from the Netherlands, China, Germany, Italy and the USA. Furthermore, they observed that research related to process mining is increasing year by year, demonstrating that it is an area that sparks growing interest and has a promising relevance for future research.

### **2.1.2. Process Discovery**

The first of the three basic types of process mining is process discovery. The purpose of process discovery is to learn through the recorded event logs by identifying patterns and traces hidden within the collection of events and to produce a process model that explains and captures the behavior seen in the data (van der Aalst, 2016; van der Aalst, 2018). As presented in Figure 2 (van der Aalst, Adriansyah, de Medeiros, et al., 2012), process discovery input should be an event log, from which constructs an accurate model of a process, without having the example of any a priori model or information. The purpose is to illustrate the actual situation instead of the one that is idealized or subjectively documented (van der Aalst, Adriansyah, de Medeiros, et al., 2012; Aguirre, Parra, & Sepúlveda, 2017). The established model is usually a process model in the Petri net or BPMN notation format, but it can also be seen from other perspectives, such as a social network (van der Aalst, Adriansyah, de Medeiros, et al., 2012). Moreover, process discovery is frequently used as a preliminary step for other practices and analysis (van der Aalst, Adriansyah, & van Dongen, 2012).

Process discovery is a demanding but simultaneously auspicious technique, especially for organizations with complex processes or aspiring to improve their digital transformation. Such companies employ process mining to find modern and distinctive ways to manage complexity and increase transparency (Reinkemeyer, 2020). Succinctly, process discovery technology enables the acceleration of processes' understandability, modeling, and redesign (dos Santos Garcia et al., 2019).

### **2.1.3. Process Conformance**

The second process mining technique is conformance checking. This procedure compares a previously constructed process model with an event log related to the same process, monitoring the discrepancies between the real process and the modeled process (van der Aalst, Adriansyah, de Medeiros, et al., 2012). Conformance checking enables the detection, localization, and explanation of deviations and quantifies and measures their severity (van der Aalst, 2016). Once again, as exemplified in Figure 2 (van der Aalst, Adriansyah, de Medeiros, et al., 2012), this technique's input is an event log and an a priori constructed model. The outcome is a diagnostic of the analysis indicating the deviations or similarities between the model and the records (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

Therefore, conformance checking is used to check if the reality recorded in the event data conforms to the process model and the other way round (van der Aalst, Adriansyah, & van Dongen, 2012). However, it has other relevant applications, such as controlling if the process is in accordance with the regulation and the documented procedures and discovering possible acts of fraud or non-conformity (Aguirre et al., 2017).

Concerning the output of this technique, it is necessary to use quality criteria to measure the level of compliance. The four evaluation metrics most used in the literature are fitness, simplicity, generalization and precision. Fitness refers to the ratio of traces that a discovered process model can reproduce from the event log. In other words, it quantifies how much of the behavior observed in the event log can be replicated by the process model. Therefore, a fitness of 1.0 means that the process model is able to reproduce all traces in the event log covering most of the behavior observed (Dunzer, Stierle, Matzner, & Baier, 2019). Simplicity expresses the complexity of a model, considering its readability and interpretation (Real, Pimentel, de Oliveira, Braga, & Stiubiener, 2020). It represents the minimal structure reflecting the behavior observed (Vidal, Vázquez-Barreiros, Lama, & Mucientes, 2016). Precision measures how well the process model represents the behavior captured by the event log; a precise model avoids underfitting (Munoz-Gama & Carmona, 2011). Conversely, generalization indicates the degree of abstraction, determining whether the process model comprehensively characterizes the process paths or only represents the captured behavior, which would mean it is overfitting by being too specific (Dunzer et al., 2019).

The work of Ly, Maggi, Montali, Rinderle-Ma and van der Aalst (2013) identified some relevant functionalities in the process conformance technique: time metrics, case data, resource information, allow explicit and implicit approaches with multiple events and different orders, support activity lifecycle, multiple-instances constraints, detect and resolve compliance violations, identify and manage violations, disclose the root cause of violations and quantify the degree of compliance or non-compliance (dos Santos Garcia et al., 2019).

#### **2.1.4. Process Enhancement**

The third type of process mining is process enhancement, also identified as process extension. The purpose of this technique is to improve or extend a priori process model based on information acquired from event logs of the actual process (van der Aalst, 2016). Accordingly, process enhancement input should be an event and an existing model. The expected output is an expanded or improved model, as seen in Figure 2 (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

Process enhancement supports two types of enhancement. The first is repair, which consists of adapting the model in order to replicate the reality more accurately. The second type corresponds to the process extension by adding a new perspective into the model and cross-correlating it with the event log (van der Aalst, 2016).

Therefore, this last main technique aims at achieving process optimization. The use of timestamps in the event log is an illustrative case of an improvement by displaying bottlenecks, service levels and execution time cycles (van der Aalst, 2016; van der Aalst, Adriansyah, de Medeiros, et al., 2012).

According to Santos Garcia et al. (2019), process enhancement can comprise the following activities and approaches: extending models, predictive techniques, organizational perspective, concept drift, decision mining and resource allocation.

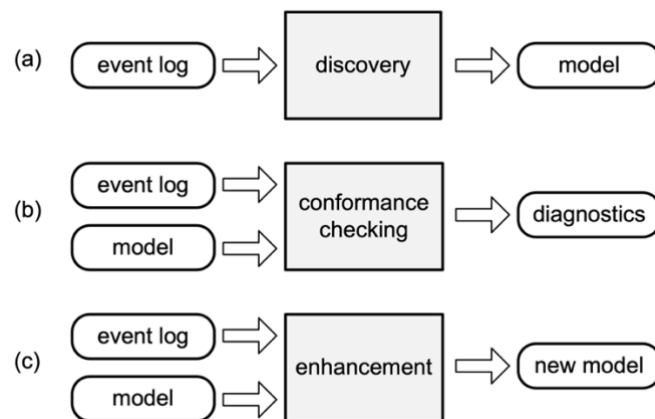


Figure 2 - The three basic types of process mining explained in terms of input and output: (a) discovery, (b) conformance checking, and (c) enhancement (van der Aalst, Adriansyah, de Medeiros, et al., 2012)

### 2.1.5. Tools & Algorithms

Despite the relatively brief existence of process mining, there is already a remarkable variety of software tools allowing the application of process mining techniques. The three most used and established are the ProM Framework, Disco and RapidProM (Rojas, Munoz-Gama, Sepúlveda, & Capurro, 2016). ProM is an open-source tool created to mitigate the problem from each tool using its own format for reading and storing, which results in different outcomes and makes it difficult to compare the results (van Dongen, de Medeiros, Verbeek, Weijters, & van der Aalst, 2005). ProM is an extensible and flexible framework that supports several process mining techniques organized as plug-ins (ProM Tools, n.d.). ProM integrates several existing functionalities, supporting many formats and languages (van Dongen et al., 2005), making it the most widely applied and recognized tool. The second tool is Disco, its main advantage is to focus on a pleasant visual interface with friendly-user functionality that allows filtering options in the event logs (Rojas et al., 2016). The third is RapidProM, which consists of developing process mining workflows resulting from the combination of RapidMiner, enabling the definition and execution of data analysis solutions, with the ProM framework and its implemented plug-ins (RapidProM, n.d.; Rojas et al., 2016). ProM is considered the most popular of the considered tools, even though Disco is more accessible and practical (Corallo, Lazoi, & Striani, 2020).

Concerning the algorithms used, the most acknowledged are Heuristics Miner and Fuzzy Miner. Heuristics Miner is a practical discovery algorithm that is able to manage noise, and its primary purpose is to produce process models by expressing its main behavior and finding the control-flow perspective (Rojas et al., 2016; Weijters, van Der Aalst, & De Medeiros, 2006). Fuzzy Miner is a configurable discovery algorithm that supports process models' development at different levels of detail by cleaning large amounts of complex behavior and dealing with unstructured event logs (Rojas et al., 2016; Günther & van der Aalst, 2007). However, more recently, there was an increasing use of new

algorithms, such as Alpha Miner, Inductive Miner and Genetic Miner (dos Santos Garcia et al., 2019; Corallo et al., 2020).

Regarding conformance checking techniques, the two techniques that stand out in the literature are Linear Temporal Logic (LTL) Checker and the Conformance Checker. The first tests if the logs satisfy some LTL formula. It compares a set of requirements with the LTL temporal logic instead of comparing a model with the log. Conformance checker requires a model and an event log to replay it within a Petri net model and simultaneously gathers diagnostic information (Bogarín et al., 2018).

These algorithms are available in tools such as ProM. Besides supporting the development of models for a wide variety of processes, they also contribute to the faster development of process models and have a simple modeling language and approach.

### 2.1.6. Guiding Principles

The “Process Mining Manifesto” (van der Aalst, Adriansyah, de Medeiros, et al., 2012) established six guiding principles to advise users, researchers and analysts about crucial aspects of the new technology and prevent vital mistakes.

- 1) **Event Data Should Be Treated as First-Class Citizens.** As previously said, events are the beginning of process mining, and the quality of the results is inherently related to the data quality. Therefore, in order for the event log to have the highest possible quality, they must have the following characteristics: be trustworthy, ensuring that the events really happened and that the attributes incorporated in the event log are correct; be complete, having all the events and attributes necessary; have explicit and clear semantics; be safe, guaranteeing privacy and security concerns when recording the event logs.
- 2) **Log Extraction Should Be Driven by Questions.** In order to obtain significant and purposeful event data, it is essential to consider concrete questions.
- 3) **Concurrency, Choice and Other Basic Control-Flow Constructs Should Be Supported.** Process mining techniques should support basic workflow patterns.
- 4) **Events Should Be Related to Model Elements.** The relation between the elements in the model and event logs has an essential role in process mining techniques, especially conformance checking and process enhancement. It depends on the relationship to discover the dissimilarity between the recorded events and the model, and it is also necessary to enhance the model with additional information.
- 5) **Models Should Be Treated as Purposeful Abstractions of Reality.** The model should be a representation of reality, a purposeful abstraction with multiple possible views captured by the event log.
- 6) **Process Mining Should Be a Continuous Process.** Processes, regardless of the industry or organization, are dynamic and changeable. Therefore, it should be an adaptable activity and not a one-time occurrence since the goal is to breathe life into process models.

### 2.1.7. Challenges

In addition to the guiding principles, the “Process Mining Manifesto” (van der Aalst, Adriansyah, de Medeiros, et al., 2012) also identified eleven main challenges to overcome when working with process mining.

- 1) **Finding, Merging, and Cleaning Event Data.** As previously mentioned, data quality and its suitability are essential factors when using process mining. Therefore, in the extraction process, it is crucial to beware of the following problems: event data distributed by various sources, data missing or incomplete, event logs with outliers or noise and logs with events at different levels of granularity. Additionally, it is necessary to consider that event data is often object-centric instead of process-centric.
- 2) **Dealing with Complex Event Logs Having Diverse Characteristics.** Event logs may have a wide range of characteristics: some can be large and with over-detailed data, making it difficult to manage, whereas others may not have all the necessary data to obtain trustworthy conclusions.
- 3) **Creating Representative Benchmarks.** Due to the recent emerging of the technique, there is still a lack of appropriate benchmarks. Good benchmarks should contain example datasets and adequate quality criteria to compare and enhance systems, programs and algorithms.
- 4) **Dealing with Concept Drift.** It is essential to be aware that the process under analysis can change due to its dynamic nature. Therefore, it is of equal importance to know how to manage such impactful changes.
- 5) **Improving the Representational Bias Used for Process Discovery.** Representational bias happens due to the use of particular languages for model construction; however, it is necessary to be conscious and careful to guarantee the quality of process discovery results.
- 6) **Balancing Between Quality Criteria such as Fitness, Simplicity, Precision, and Generalization.** The balance of these four quality dimensions can be challenging since they might conflict.
- 7) **Cross-Organizational Mining.** Commonly, process mining is explored within a particular organization. However, one of the two cross-organizational process mining cases can happen: organizations working together on process instances or different organizations working in similar processes sharing knowledge or tools. In addition to managing these two types of cross-organizational, it is also necessary to consider privacy issues emerging from the cross-organizational approach.
- 8) **Providing Operational Support.** In its beginning, process mining was more directed towards the analysis of historical data. However, with the increasing access to real-time data and the greater capacity for computing power, there is now a greater use of online operational support, which can be classified into three techniques: detect, predict, and recommend.
- 9) **Combining Process Mining with Other Types of Analysis.** Bringing together process mining with other analysis techniques might be difficult, but it will originate valuable insights if it is successful.
- 10) **Improving Usability for Nonexperts.** Hiding the complex process mining algorithms through user-friendly interfaces to achieve the goal of having process models used on a daily basis with automatic parameters and suggestions so that end-users can interact with the results.
- 11) **Improving Understandability for Nonexperts.** The end-user may have difficulty understanding the output and eventually make inappropriate conclusions. To overcome this, the results should be clear and with an appropriate representation. Moreover, it should include the trustworthiness of the results.

Recently, Wil van der Aalst (2020) revised its previous challenges list, adapting it according to the evolution that process mining has undergone in the last eight years, and proposed the following challenges.

- 1) **Bridging the gap between process modeling and process mining.** Finding a balance between a hand-made model that describes idealized behavior and a discovered model strictly based on data is necessary in order to mitigate the lack of reliability originated from process modeling and the inadequacy of process mining in the absence of data. A possible method of converging the two is through hybrid process models.
- 2) **Incorporating stochastic information in process models to improve conformance checking and prediction.** An excellent example of this is the critical role of the frequencies of activities, which identify highly frequented process variants and represent estimates for probabilities. Moreover, it is essential to the precision of process models and for its prediction and comparison.
- 3) **Process Mining for multiple processes using different case notions.** Process mining assumes that each event corresponds to a particular case and each case to a singular process. However, the reality is more complex, and one event may refer to various cases. With the development of process mining techniques, it may be possible to integrate process models with multiple perspectives.
- 4) **Dealing with uncertain and continuous event data.** With the increasing use of process mining in the most diverse scopes and the consequent collection of data derived from different environments, it became evident the need for process mining to beware that each event may be incorrect or not happened as described. Besides, some events depend on their continuous variables to represent the actual meaning of the event.
- 5) **Comparative Process Mining to identify differences between process variants over time.** Processes are dynamic and change over time. Besides, the same process may have a different variable, such as being performed at different locations. Hence, the importance of comparing the corresponding process variants in the analysis.
- 6) **Causality-aware Process Mining ensuring correct and fair conclusions.** Process mining techniques can help uncover performance and compliance problems through bottleneck analysis and conformance checking results. Additionally, integrating data mining and machine learning will expand the knowledge and uncover root causes for both problems. Nevertheless, correlation does not mean causation, which leads to incorrect or unfair conclusions. Hence, the necessity of applying explicit causal models to guide root-cause analysis in process mining.
- 7) **Confidentiality-aware Process Mining to avoid unintentionally leaking sensitive information.** Preserve confidentiality and data privacy are major concerns in process mining, especially after the implementation of the EU General Data Protection Regulation. However, event data is susceptible since the order of activities or a few timestamped events are enough for identification. Therefore, there is research currently devoted to anonymization and encryption techniques.

## 2.2. EDUCATION

### 2.2.1. Concepts

Education is a fundamental necessity and a powerful instrument for the population by supporting each individual in its development as persons, citizens and professionals (OECD, 2019). Additionally, it contributes to the evolution of society, the development of the economy and the progress of all countries. Education is, therefore, considered the primary tool for world development, contributing to a more sustainable and civilized planet (ECSSR, 2019).

Nevertheless, education encompasses a high level of complexity. Whether because it is intrinsically dependent on the environment and the society surrounding it, which requires a constant need of adaption in order to reflect social and cultural developments. Alternatively, because it is also necessary to consider the constant changes in perspectives and theories concerning education, which was accentuated with the current digitalization of the modern world and the speed of change due to the evolution and increasing use of information communication technologies along with a wide variety of other innovations (Rix & Twining, 2007). Besides, the complexity in education is also due to the fact that it is about and for human beings, each with different needs and particular experiences, thus involving numerous factors and characteristics for each specific circumstance and personality (Turner, 2009).

In the interest of clarifying and facilitating the understanding of this broad term, there are three well-established concepts that intend to distinguish three very different forms of education:

- 1) Formal education is the institutionalized, premeditated and programmed education. The national education authorities recognize the agenda and program, whether private or public organizations, in order to represent a country's educational system. Therefore, formal education occurs predominantly in educational institutions and involves a structured educational arrangement, characterized by a student-teacher relationship in a full-time regime intended to acquire knowledge, skills and competences, which will result in qualifications and certifications (OECD, 2010; UIS, 2012).
- 2) Non-formal education, similarly to the formal it is institutionalized, premeditated and programmed. However, it functions as a complement or alternative to formal education, usually in the form of workshops, short courses or seminars. Non-formal education does not depend on the qualification recognized by education authorities. Non-formal education comprises activities ranging from work skills and social or cultural development to personal development activities (UIS, 2012).
- 3) Informal education represents all the forms of learning that are not institutionalized but are intended or deliberate. Informal learning encompasses the knowledge obtained through our daily activities, whether related to work, family or local community, based on self-learning, family-directed or socially. It is, therefore, considerably less organized and structured than the previous two types (UIS, 2012). It is often recognized simply as experience, considering that people throughout their lives are inevitably exposed to learning situations (OECD, 2010).

Education systems differ widely between countries, both in terms of curriculum and structure, complicating a reliable comparison and identification of the different educational systems. As a means to mitigate this obstacle and have international education data, all based on the same criteria, the United Nations Educational, Scientific and Cultural Organization (UNESCO) established the International Standard Classification of Education (ISCED 2011), which defines a classification possible to be used for all countries. It is based on well-defined criteria that allow the allocation of educational programs to the following levels (UIS, 2012):

- ISCED 0: Early childhood education
- ISCED 1: Primary education
- ISCED 2: Lower secondary education
- ISCED 3: Upper secondary education

- ISCED 4: Post-secondary non-tertiary education
- ISCED 5: Short-cycle tertiary education
- ISCED 6: Bachelor's or equivalent level
- ISCED 7: Master's or equivalent level
- ISCED 8: Doctoral or equivalent level

Furthermore, UNESCO developed the ISCED Fields of Education and Training (ISCED-F 2013), which classifies the main fields of education broadly (UIS, 2014):

- 00 Generic programmes and qualifications
- 01 Education
- 02 Arts and humanities
- 03 Social sciences, journalism and information
- 04 Business, administration and law
- 05 Natural sciences, mathematics and statistics
- 06 Information and Communication Technologies
- 07 Engineering, manufacturing and construction
- 08 Agriculture, forestry, fisheries and veterinary
- 09 Health and welfare
- 10 Services

### **2.2.2. Educational Processes**

A simplified way of approaching education can be expressed through the two main participants of the learning activity, teachers and students. Teachers provide the appropriate resources and services to ensure student success. As the essential part of any learning activities, students interact with other participants and the system itself, which means courses, lectures and exams (Bogarín et al., 2018).

Nevertheless, education relies on and comprises several processes. The most important processes in education are Learning, Teaching, Assessment, Administration and Resource Management (IBE, 2012).

Learning is the core process of the whole educational system, considering it is the reason for its creation. It should serve the purpose of facilitating the learning process, which is a complex and lifelong process, by structuring and controlling (IBE, 2012).

One way to simplify and summarize the act of learning is by dividing it into the following three main processes: acquisition, which refers to the acquirement of new information, which generally works as a substitute and modification of the information already acquired or as an extension of some knowledge or understanding already present; transformation, it is the process in which a natural manipulation of knowledge takes place that allows to deal with the information previously acquired in a way that goes beyond it, whether extrapolating it or converting it into another form; and evaluation, its purpose is to analyze if the information was manipulated adequately, that is, if the generalizations made are convenient or if the extrapolations are appropriate (Bruner, 1977; IBE, 2013).

Teaching is the central and most direct process to support learning and allow the learners to acquire the expected skills or the knowledge passed through the given information (IBE, 2012). There are various teaching methods, from a more didactic approach that mainly refers to the lecturing, being,

consequently, centered on a teacher-student relationship and is content-oriented, since the role of learners is solely to receive the information transmitted by the teacher. To a more supporting process, which is learner-centered and aims at building on the knowledge and understanding that the student already has. Another approach focuses on students' cognitive development and critical thinking in order to have greater awareness and control of thinking and learning (IBE, 2013).

Assessment is commonly understood as the process that quantifies, measures and judges the student's progress according to specific and defined quality criteria (IBE, 2013). However, the assessment covers and supports different areas of the educational process, depending on its purpose, form or area of assessment. It is, therefore, crucial to distinguish between two particular forms of assessment, the assessment for learning and the assessment of learning (IBE, 2012). The former represents the assessment of the learner's progress and accomplishments, consequently, its primary aim is to support and improve the learning process by adapting the educational process to the student's needs. Additionally, it involves informing students of their strengths and weaknesses and providing support and feedback to mitigate their difficulties (IBE, 2013). The latter is related to classical measurement, the performance of tests and examinations, which provides information about the student's achievements and what has been learned in a certain period (IBE, 2013). Nonetheless, it only provides insights from analyzes of the final results (Tóth, Rölke, Goldhammer, & Barkow, 2017). Furthermore, it is usually associated with its usage for promotion or graduation. Besides, assessment of learning plays an essential role in monitoring and controlling the achievement of the educational system itself, crucial to assess the quality of education and the effectiveness of the implemented curriculum (IBE, 2012).

Administration and Resource Management play a crucial role in the whole educational system. They guarantee that students' and teachers' necessities are met, and they are accountable for recording, checking and analyzing students' data. Administration includes, for example, student registration processes which deal with every requirement and step necessary to the academic or course registration. In comparison, Resource Management can include the management of the software repositories and software development processes. Source-code management systems, mail archives, document repositories, version control systems and bug trackers are examples of such repositories (Bogarín et al., 2018).

### **2.3. PROCESS MINING IN EDUCATION**

The principal role of education systems is to improve learners' development, growth and learning outcomes (Jornitz & Engel, 2021). The normalization of digital data in the world and, mainly, in education has made its use essential for optimizing the quality and effectiveness of the educational system (Selwyn, 2015; Jornitz & Engel, 2021). Nowadays, digital data has taken on such significance that the information obtained from the accumulation and flow of data is considered necessary for the entire understanding of contemporary education (Selwyn, 2015).

This section introduces the systematic literature review on the application of process mining in the education domain, which intends to collect, analyze and summarize the most relevant existing literature findings. The execution of a systematic literature review consists of a series of phases, such as identifying the potentially relevant records, screening the previously identified records, applying eligibility criteria of the included studies and a detailed analysis of the final selection (Liberati et al., 2009).

The systematic review will follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and its corresponding flow diagram (Moher, Liberati, Tetzlaff, Altman, & The, 2009). PRISMA is characterized by being a transparent and complete reporting tool, contributing to the discovery of relevant and appropriate research (Liberati et al., 2009).

### 2.3.1. Search Strategy

The search process occurred in four electronic databases: IEEE Xplore, Science Direct, Springer Link and Google Scholar. It was executed in April 2021, with the aim of reviewing scientific and technological literature from conferences and journals. It followed the following phases, illustrated in the PRISMA flow diagram in Figure 3 and described below.

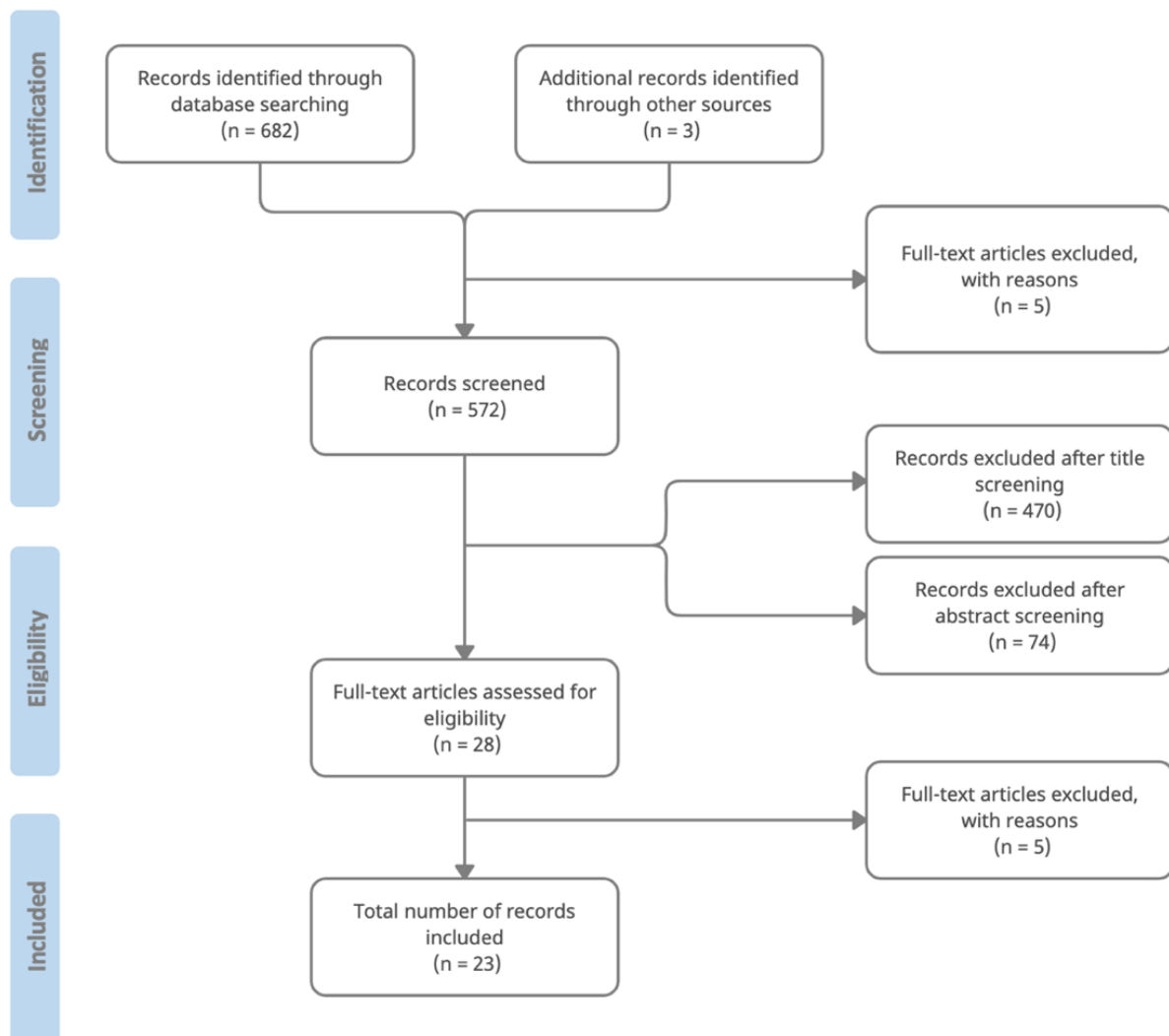


Figure 3 - PRISMA flow diagram

In phase 1 was applied the search string, developed with Boolean operators and represented in the following Figure 4, in the indicated repositories to research articles and papers issued between 2015 and April 2021. This first step of the research obtained 682 results, plus three additional records identified through other sources, which, in this case, means obtained through other articles.

| Digital Database      | Search String                                                                                                                                                                                                                   |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>IEEE Xplore</i>    | ("education" OR "educational" OR "learning" OR "learning analytics" OR "learning management" OR "education domain" OR "school") AND ("process mining" OR "process discovery" OR "conformance checking" OR "performance mining") |
| <i>Science Direct</i> | ("education" OR "educational" OR "learning" OR "learning analytics" OR "learning management" OR "education domain" OR "school") AND ("process mining" OR "process discovery")                                                   |
| <i>Springer Link</i>  | "process mining" AND (education OR educational OR learning OR school)                                                                                                                                                           |
| <i>Google Scholar</i> | allintitle: "process mining" education OR educational OR learning OR "learning analytics" OR "learning management" OR "education domain" OR school                                                                              |

Figure 4 - Search Strings

The second phase includes the screening of the articles initially found, beginning with the elimination of duplicate records, articles not published in English and records with no access available, resulting in the exclusion of 113 records. Which was followed by a selection based on the titles and abstracts screening, resulting in the exclusion of position papers, workshops, prefaces, announcements, call for papers, corrections and, mainly, articles not explicitly related to the topic and the research questions. Consequently, with this phase, the papers' selection was narrowed to 28 publications, being 470 records excluded from the title screening and 74 from the abstract and keyword screening.

In phase 3 belongs the detailed revision and analysis of the entire text of each paper by checking for eligibility, which led to the elimination of five articles. Remaining, therefore, 23 records in the fourth and last phase, which represent the final selection of articles included in this study for analysis and discussion, which follows in the next section and is summarized in Table 1.

Table 1 - Records Summary

| Paper | Title                                                                                                   | Reference                                  | Application                                                                                                               | Data dimension                                                                                                                                      |
|-------|---------------------------------------------------------------------------------------------------------|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| S1    | Using process mining to analyze students' quiz-taking behavior patterns in a learning management system | (Juhaňák, Zounek, & Rohlíková, 2019)       | Identify specific types of interaction sequences that enhance the knowledge about students' quiz-taking strategies in LMS | LMS: Moodle; Five quizzes assigned in five different courses in Moodle at the Faculty of Arts, Masaryk University, Brno (Czech Republic)            |
| S2    | Process mining for self-regulated learning assessment in e-learning                                     | (Cerezo, Bogarín, Esteban, & Romero, 2020) | Discover students' self-regulated learning processes during an e-Learning course by using Process Mining Techniques       | LMS: Moodle; e-learning; interaction traces from 101 university students; course given on the Moodle 2.0 platform; University in the North of Spain |

| Paper | Title                                                                                                                                           | Reference                                                      | Application                                                                                                                                                                                                                       | Data dimension                                                                                                                        |
|-------|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| S3    | Educational process mining: a tutorial and case study using Moodle data sets                                                                    | (Cristóbal Romero, Cerezo, Bogarín, & Sánchez-Santillán, 2016) | Discover the most specific browsing behaviors when using only the clustered data rather than the whole dataset                                                                                                                    | LMS: Moodle; e-learning; Moodle 2.0 course used by 84 undergraduate students; Psychology degree program, University in northern Spain |
| S4    | The design of exploratory and preprocessing of event log data in online learning activities based on Moodle LMS for process mining              | (Aulia & Waspada, 2019)                                        | Build an application to contribute to the Preprocessing and Exploratory Data Analysis stages of Moodle event log data                                                                                                             | LMS: Moodle; online learning data                                                                                                     |
| S5    | Implementation of Alpha Miner algorithm in process mining application development for online learning activities based on Moodle event log data | (Nafasa, Waspada, Bahtiar, & Wibowo, 2019)                     | Develop the application to discover process model and integrate the Moodle event log data with process mining; facilitate the preprocessing tools                                                                                 | LMS: Moodle; online learning process data from a course downloaded directly from the Moodle system                                    |
| S6    | Discovering time management strategies in learning processes using process mining techniques                                                    | (Uzir et al., 2019)                                            | Propose an analytic learning methodology to analyze the link between self-regulated learning (SRL) and time management practices                                                                                                  | LMS: Moodle; first-year students (N = 241) enrolled in blended learning of a health science course                                    |
| S7    | Educational process mining for verifying student learning paths in an introductory programming course                                           | (Real, Pimentel, de Oliveira, Braga, & Stiubiener, 2020)       | Present the application results of PM techniques to verify the students learning paths in an introductory programming course                                                                                                      | LMS: Moodle; blended learning; 24 605 events collected from 73 undergraduate students                                                 |
| S8    | Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs                                                         | (Álvarez, Fabra, Hernández, & Ezpeleta, 2016)                  | Use of an LTL-based model checking tool to analyze Moodle logs corresponding to a degree subject                                                                                                                                  | LMS: Moodle; 107 students enrolled in a Moodle course                                                                                 |
| S9    | Using Process Mining Techniques to Discover Student's Activities, Navigation Paths, and Behavior in LMS Moodle                                  | (Dolak, 2019)                                                  | Discover student's activities, navigation paths, and behavior; propose a two-step clustering approach to extract the best training paths                                                                                          | LMS: Moodle; Online course; 701 students from Silesian University in Opava, School of Business Administration in Karvina              |
| S10   | Trace-SRL: A framework for analysis of microlevel processes of Self-Regulated Learning from trace data                                          | (Saint, Whitelock-Wainwright, Gašević, & Pardo, 2020)          | Present the "Trace-SRL" framework that comprises: 1) clustering of learner types; 2) use of microlevel processing to transform raw trace data into SRL processes; 3) use of process mining to explore the generated SRL processes | LMS; trace data from online activities, 300 computer engineering undergraduate students                                               |
| S11   | Detecting learning strategies through process mining                                                                                            | (Saint, Gašević, & Pardo, 2018)                                | Explore the application of process mining to identify strategic and tactical learner behaviors; performed intra- and inter-strategy analysis                                                                                      | LMS; online activities; 300 computer engineering undergraduate students                                                               |

| Paper | Title                                                                                                                              | Reference                                                                    | Application                                                                                                                                                                                      | Data dimension                                                                                                                                                                                                       |
|-------|------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S12   | Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses          | (Maldonado-Mahauad, Pérez-Sanagustín, Kizilcec, Morales, & Munoz-Gama, 2018) | Apply process mining from the SRL-perspective, identify three clusters of learners and six distinct interaction sequence patterns                                                                | 3458 learners across three Massive Open Online Courses (MOOCs)                                                                                                                                                       |
| S13   | Learning analytics on Coursera event data: A process mining approach                                                               | (Mukala, Buijs, Leemans, & van der Aalst, 2015)                              | Use process mining techniques to trace and analyze students' learning habits based on MOOC data and relate them with students' performance                                                       | MOOC data, Coursera platform; dataset contains information about students, clickstream and course structure                                                                                                          |
| S14   | Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering | (van den Beemt, Buijs, & van der Aalst, 2018)                                | Explore students' activities in a MOOC from the perspective of personal constructivism, combining learning behavior and learning progress                                                        | MOOC data, Coursera platform; clickstream related with video lectures or quizzes submission                                                                                                                          |
| S15   | On predicting academic performance with process mining in learning analytics                                                       | (Umer, Susnjak, Mathrani, & Suriadi, 2017)                                   | Propose a process mining approach to make early predictions to improve students' learning experience in MOOCs; investigate the combination of machine learning techniques with process mining    | MOOC data, Coursera platform; 3000 students; dataset contains assessments grades, solution submission time, video lecture interaction log, participant's demographic information, time spent weekly and final grades |
| S16   | Business process reporting using process mining, analytic workflows and process cubes: a case study in education                   | (Bolt, de Leoni, van der Aalst, & Gorissen, 2017)                            | Explore the combination of process mining, process cubes and analytic workflows to overcome the limitation of each technique and obtain reliable and valuable insights into the student behavior | Eindhoven University of Technology IT systems; dataset contains video lecture views, course grades and personal information of students                                                                              |
| S17   | Process-mining enabled feedback: "Tell me what I did wrong" vs. "tell me how to do it right"                                       | (Sedrakyan, De Weerd, & Snoeck, 2016)                                        | Explore various learning behavior patterns linked with learning outcomes; analyze the complex problem-solving process                                                                            | Data of 36 cases: event logs of students' group works                                                                                                                                                                |
| S18   | Educational process mining: New possibilities for understanding students' problem-solving skills                                   | (Tóth, Rölke, Goldhammer, & Barkow, 2017)                                    | Describe how to extract process-related information from event logs and how to use these data in online problem-solving assessments                                                              | Students' interactions with the assessment system                                                                                                                                                                    |
| S19   | Recompiling learning processes from event logs                                                                                     | (Vidal, Vázquez-Barreiros, Lama, & Mucientes, 2016)                          | Present an approach to extract the learning flow structure, to obtain the underlying rules that control students' adaptive learning and to reuse units of learning                               | Virtual Learning Environment event log                                                                                                                                                                               |

| Paper | Title                                                                                                                         | Reference                                          | Application                                                                                                                                          | Data dimension                                                                                                 |
|-------|-------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| S20   | Process mining application for discovering student learning paths                                                             | (Intayoad, Kamyod, & Temdee, 2018)                 | Extract learning process models from student groups with different learning background to enhance personalized learning                              | Two data sources: online learning database and web usage data                                                  |
| S21   | Influence of student diversity on educational trajectories in engineering high-failure rate courses that lead to late dropout | (Salazar-Fernandez, Sepúlveda, & Munoz-Gama, 2019) | Analyze educational trajectories using process mining techniques to explain the relationship between a sequence of academic results and late dropout | Academic records; 794 engineering students at Universidad Austral de Chile                                     |
| S22   | Process Mining in the Education Domain                                                                                        | (Cairns, Gueni, Fhima, et al., 2015)               | Investigate the potential, challenges and feasibility of the educational process mining in the field of professional training                        | Real-world training databases from a worldwide consulting company with 6000 employees; 2272 training courses   |
| S23   | Analyzing and improving educational process models using process mining techniques                                            | (Cairns, Gueni, Assu, Joubert, & Khelifa, 2015)    | Demonstrate the applicability of process mining to monitor and analyze educational processes in the field of professional training                   | Real-world training databases from a worldwide consulting company with 6000 employees; 16 260 training courses |

Figure 5 illustrates the distribution by year of all papers found and the 23 included in the study. Through the line referring to all papers, it is possible to notice an increase in research on the subject in recent years that is not noticeable with the few papers included in the study.

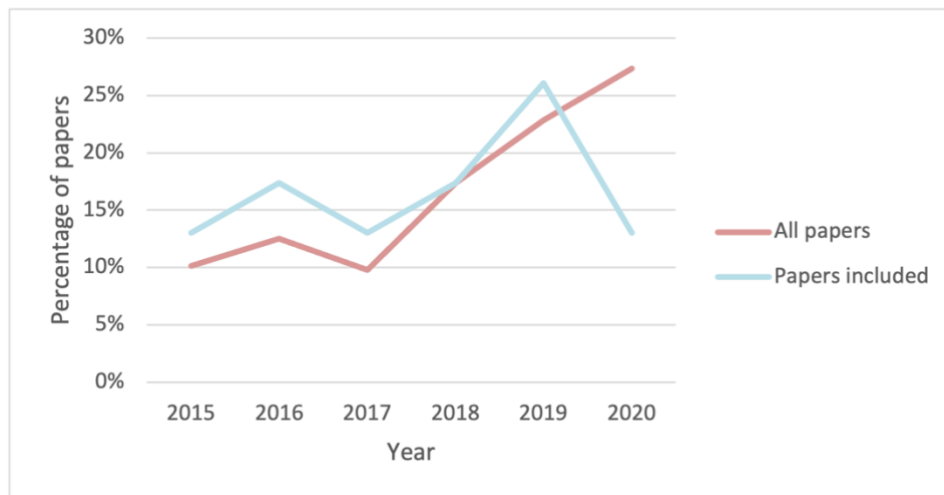


Figure 5 - Percentage of papers published per year

### 2.3.2. Analysis and Discussion

The implementation of process mining in the field of education, applied to educational data, is called Educational Process Mining (EPM). EPM supports the extraction of process-related knowledge from event logs captured by an educational system (Cristobal Romero & Ventura, 2013). EPM has the means for building complete and compact educational process models that reproduce the observed behavior, check the conformance of the modeling behavior with the observed behavior, extend the information to make tacit knowledge explicit (Bogarín et al., 2018).

Therefore, EPM application can contribute to a better comprehension of the underlying educational process by generating feedback to students, teachers, researchers and institutions, and consequently, giving recommendations and advice to students or improving the management of learning objects (Bogarín et al., 2014). Moreover, EPM supports the exploration of students' learning paths, their profile and behavior and, additionally, the performance of conformance analysis and model extension.

Furthermore, the potentials of EPM can be employed in innumerable forms of education, from formal and presential educational systems (papers: S8, S10, S11, S17 and S21) to a variety of distant education programs, such as online learning courses (papers: S2-S5, S9, S18 and S20) and MOOCs (papers: S12-S15), as well as blended learning courses (papers: S1, S6, S7 and S16). Correspondingly, it can be incorporated within Learning Management Systems (LMS), such as Moodle (papers: S1-S11) and other Virtual Learning Environments (VLE) (S19).

The papers S3, S7, S8, S9, S13 and S14 represent various ways in different environments of using process mining to discover students' behavior paths and learning processes and relate them to the results obtained, usually expressed through final marks.

The studies S3, S7 and S8 based their works on log files stored by LMS Moodle. In the first two papers, the researchers Cristóbal Romero et al. (2016) and Real et al. (2020) segmented the students according to passed and failed results to obtain models with high fitness and performance levels. Both studies obtained similar results: the learning paths taken by approved students are different from the paths of failed students concerning the activities undertaken in the course. The third study also analyzed the relationship between students' behavior and their final grades. However, Álvarez et al. (2016) have extended their research by applying a model checking tool. The findings revealed that the students' behavior did not fit the teachers' plan.

The studies S13 (Mukala et al., 2015) and S14 (van den Beemt et al., 2018) have an approach similar to the previous study, whether about using process mining and clustering or even concerning conformance checking techniques. However, these two studies focus on the learning process based on MOOC data. Both studies agree that students who regularly watch videos and in a batch are more likely to have positive results and be successful.

The study S9 by Dolak (2019) provides another example of the utility of EPM over Moodle event logs. The authors developed a detailed study of learners' behavior based on how active or inactive they were in their learning process.

The research of the learners' behavior also includes the evaluation and assessment activity. Studies S1, S17 and S18 cover this specific part of the students' learning assessment extensively. The Juhaňák et al. (2019) in the study (S1) provided a comprehensive analysis of the students' interactions while

taking online quizzes in Moodle, resulting, therefore, in the discovery of the sequence of activities. S17 and S18 explored the behavior patterns in problem-solving processes. Nevertheless, they have adopted two very different approaches on the subject. The former is a study by Sedrakyan et al. (2016) that aimed to discover perspectives of process-oriented feedback based on process mining analysis on learning processes to replace the popular outcome feedback, only available at the end of a task. The latter (Tóth et al., 2017) describes how test-taking log data can be incorporated in educational assessment with process mining methods, including visualization, clustering and classification methods.

Another use of EPM for analyzing data from LMS, particularly Moodle, is by studying the Self-Regulated Learning (SRL) processes (papers: S2, S6, S10, S11 and S12). Self-Regulated learning represents a process and an aggregation of thoughts, feelings and actions originated by the students' willingness to achieve their objectives and purpose (Cerezo et al., 2020).

In article S2, Cerezo et al. (2020) assessed the students' SRL skills throughout an e-learning course using EPM techniques on Moodle logs. They found that approved students did not precisely follow the instructors' suggestions but followed the behavior described in a successful self-regulated process. In a similar environment, Uzir et al. (2019), Saint et al. (2020) and Saint et al. (2018) (papers: S6, S10 and S11) proposed a methodology that combined process mining with cluster analysis to study the link between SRL and time management practices as well as the identification of strategic and tactical learner behaviors. In the last article (Saint et al., 2018), the authors developed a framework, the "Trace-SRL", to explore the use of micro-level process analysis with process mining to identify traits of SRL, categorize learners' actions and compare learning strategies with performance.

In addition, and still related to the SRL processes, much research in recent years has focused on the analysis of the SRL processes related to MOOCs platforms. For example, the paper S12 aimed to understand and investigate the relationship between theoretical SRL strategies and observed behavioral patterns of learners on MOOCs. Maldonado-Mahauad et al. (2018) identified three clusters of learners: comprehensive learners, targeting learners and sampling learners.

The studies (S15 and S16) provided critical insights into exploring process mining combined with other techniques. The paper S15 carried out by Umer et al. (2017) demonstrated that the integration of process mining techniques with traditional machine learning algorithms performed better than standard features alone as a means to monitor students' performance in MOOCs and to predict their overall performance outcome. In contrast, Bolt et al. (2017) introduced in S16 a combination of process mining with process cubes and analytic workflows. This combination was applied on large and heterogeneous event data of multidimensional nature to automate the analysis and contribute to repetition.

Some recent research, S4 and S5 (Aulia & Waspada, 2019; Nafasa et al., 2019), have addressed the issue of implementing and exploring Moodle-based process mining, especially the preprocessing phase. Both studies built an application to facilitate the preprocessing and Exploratory Data Analysis stages of the process analysis. By using these systems, teachers and researchers would be able to use EPM on the Moodle event log data more easily in an integrated environment.

As previously stated, the application of process mining in the education field is diverse and contains multiple uses. S19, S20 and S21 represent other examples of its function distinct from the articles explored so far.

The article S19 (Vidal et al., 2016) presented a new framework to simplify the reuse of courses, seminars or workshops between different virtual learning environments (VLEs), which was from where the event data was obtained and, therefore, used to extract and build learning flow models using process mining and, furthermore, to obtain the underlying rules that control students' adaptive learning through decision tree learning.

The adoption of process mining to explore the personalized learning topic received less research attention. Intayoad et al. (2018) addressed the topic in S20. They intended to improve personalized learning by extracting learning processes based on the aggregation of the students, depending on their different learning backgrounds and their final marks.

The approach brought up by Salazar-Fernandez et al. (2019) in S21 contemplated the analysis of students' learning paths using process mining aiming to understand the relationship between academic outcomes and late school dropout. Additionally, the case study seeks to identify similarities and differences among students depending on gender, income level and level of mathematical skills.

Finally, remains a particular way to use EPM, which is related to the professional training process, that was pointed out in the studies S22 and S23. The first paper, by Cairns, Gueni, Fhima, et al. (2015), performed a process model discovery which aimed at analyzing the interactions between training providers, courses units and resources involved in the students' training paths. The work also proposes a two-step clustering approach based on Key Performance Indicators and traces in the log events. Cairns, Gueni, Assu, et al. (2015), the second study researchers, analyzed professional training processes and checked their conformance with the established curriculum and prerequisites. As in the previous paper, it was also used performance indicators, but, in this case, to enhance the training process model.

Moreover, Figure 6 presents the tools that each article used in its study. Where it is visible that ProM stands out, followed by Disco.

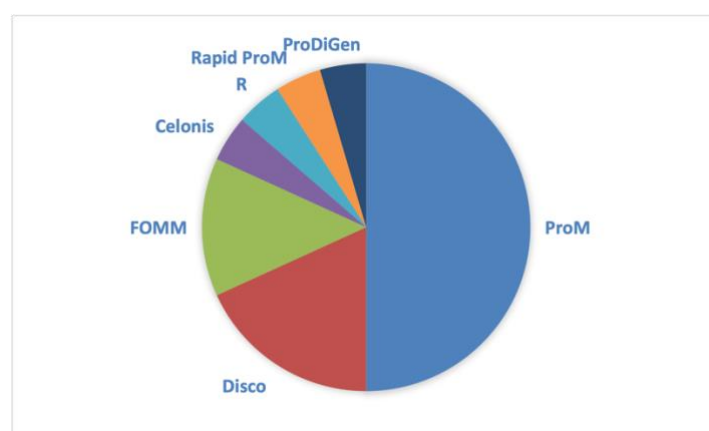


Figure 6 - Tools used in research

### 3. METHODOLOGY

From the consulted literature, it is possible to identify three main process mining methodologies, L\* life-cycle model (van der Aalst, Adriansyah, de Medeiros, et al., 2012), Process Diagnostics Method (Bozkaya, Gabriels, & Werf, 2009) and PM<sup>2</sup> Methodology (van Eck et al., 2015). The methodology used to guide the process mining exploration in this dissertation is based on the PM<sup>2</sup> Methodology (van Eck, Lu, Leemans, & van der Aalst, 2015), which offers an iterative and simple approach in its six stages. PM<sup>2</sup>, being the most recent, brought adjustments and refinements to those previously mentioned (van der Aalst, 2016); therefore, it is considered the appropriate option for this study. Furthermore, the chosen methodology is suitable for the majority of the mining techniques, for the analysis of structured and unstructured processes and for various business domains (Maldonado-Mahauad et al., 2018; Graafmans, Turetken, Poppelaars, & Fahland, 2021).

Moreover, a search in the Google Scholar database related to the usage of the three methodologies shows that 1308 results cited the paper where L\* life-cycle model was introduced. However, it refers to the Process Mining Manifesto; through further research within citing results, we have that the methodology was mentioned in only 229 results. Regarding the other two articles, the one that introduced the PDM methodology was cited in 187 results and the PM<sup>2</sup> Methodology in 236 results, which demonstrates a greater interest and research in the PM<sup>2</sup> Methodology.

#### 3.1. PM<sup>2</sup> METHODOLOGY

PM<sup>2</sup> presents instructions for the execution of process mining projects and their analysis with the aim of improving process performance or compliance. The methodology describes the possible stages in a process mining project; following the suggested steps contributes to the use of best practices and encourages the exploration of various process mining techniques as well as other analyses, resulting in advancing insights and opportunities to improve the process (van Eck et al., 2015).

PM<sup>2</sup> comprises six phases, each with its defined input and output along with specific activities to be executed (van Eck et al., 2015). Figure 7 illustrates an overview of the model developed by van Eck, Lu, Leemans and van der Aalst (2015) and a detailed description of each stage is presented below.

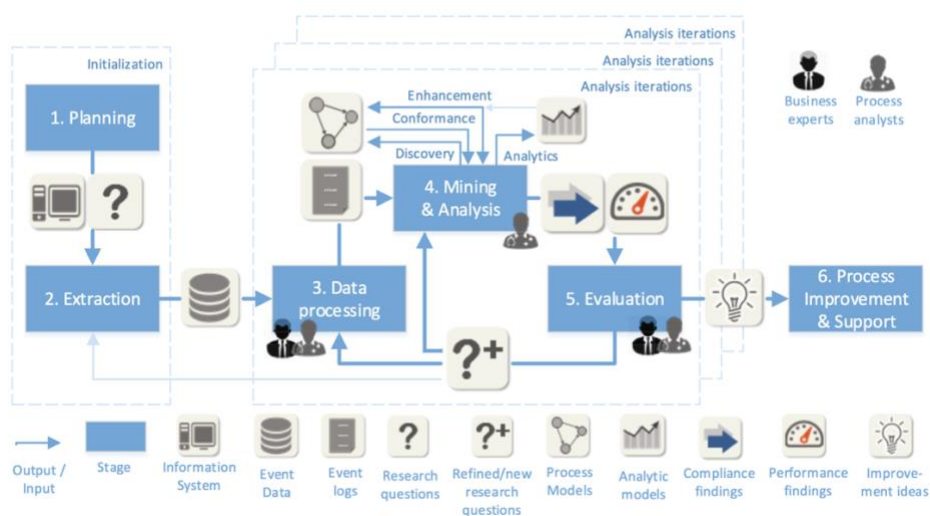


Figure 7 - An overview of the PM<sup>2</sup> methodology (van Eck et al., 2015)

1. **Planning:** this phase aims to assemble the process mining project and establish the research questions. The inputs of this level should be the organization's business processes, while the outputs should be research questions and a set of information systems. Hence, this step can include activities such as identifying the goals, translating them into research questions and selecting business processes.
2. **Extraction:** the focus of this stage is to extract event data. Its inputs are the previously defined research questions and information systems, and the output is event data. There are three activities to be considered in this step: defining the scope of the data extraction (such as granularity, period, attributes, correlation between data), extracting event data and transferring process knowledge.
3. **Data Processing:** the main idea is to generate event logs as different views of the event data and process them to be ideal for the following stage, mining and analysis. Accordingly, the inputs are event data and, in some instances, process models. The outputs are the referred event logs. Four possible activities belonging to this step are: creating views depending on the goal of the analysis, aggregating events, enriching logs with other attributes from the log or external data and, finally, filtering logs. This last activity can be further detailed into three types of filtering: attribute filtering, to remove events or traces; variance-based filtering, to group similar traces (e.g., through clustering); and compliance-based filtering, to remove traces or events that do not fit the compliance rule.
4. **Mining and Analysis:** at this stage, process mining techniques are applied. It is also intended to answer the research questions and get an insight into the process. The inputs are event logs and, once again, it can also be process models. The findings obtained through the answers to the research questions are the output of this stage. This phase is based on four activities: process discovery, conformance checking, enhancement and process analytics. The last activity comprises other complementary analysis techniques.
5. **Evaluation:** the main objective is to associate the analysis results with the project objectives' improvement ideas. The inputs required are the activities findings from the analysis stage. The outputs should be improvement ideas or new research questions obtained from the activities: diagnose and verify & validate.
6. **Process Improvement and Support:** this step aims at using the gained insights to modify the execution of the process. The inputs are the improvement ideas, and the outputs would be process modifications. Concerning the activities of this phase, the suggested ones are the implementation of improvements and support to operations.

### **3.2. RESEARCH STRATEGY**

This section describes how each stage of the PM<sup>2</sup> Methodology will be adapted and applied to this study scope.

1. **Planning:** this initialization phase can be associated with the introduction and literature review chapters of this dissertation, where the goals were defined and the research questions established, along with an extensive study of the latest literature on the application of process

mining in education. With the literature review was possible to identify the most common practices and applications, the most used techniques and tools and the existing limitations.

2. Extraction: the description of the activities related to the data extraction phase will be performed in the data collection section in the following chapter, including the log preparation, inspection and preprocessing activities.
3. Data Processing: from the previous phase, we will gain a deeper insight into the available event data, which will contribute to a better understanding of what will need to be done in the data treatment phase, in order to result in the ideal event log for the next step. In this study, data processing activities will involve aggregating events by course and semester period and filtering logs through two techniques: attribute filtering, eliminating the less common events, and variance-based filtering, through clustering. The detailed description of the aforementioned will belong to section 4.3. Data Processing.
4. Mining and Analysis: the analysis of the process mining project will be described in detail in section 4.4. Process Mining and Analysis, along with the display of the most significant process models.
5. Evaluation: the evaluation will be carried out in section 4.5. Evaluation, where the method previously applied will be used on a data set where the students who failed are not distinguishable, in order to test whether the method will be able to predict which students are most likely to be at risk of failing or not being successful.
6. Process Improvement and Support: this phase will be addressed in section 4.6. Discussion, which will explain how to approach and interpret the results, in the interest of anticipating and acting in advance in the most problematic cases or students at risk of failing, along with some recommendations to be taken into account.

## 4. STUDY

### 4.1. OVERVIEW

This section presents an overview consisting of knowledge acquired in the previous chapters and that we consider essential to emphasize for the development of this chapter. This chapter aims to extract information from event data extracted from a LMS and obtain insights. The extracted event log comprises data associated with a single process. Accordingly, a process consists of cases, and each case contains events. Each event refers to an activity performed by a particular resource with a timestamp, and therefore, events are ordered (van der Aalst, 2016). The system records the interactions to some extent in an event log; consequently, it only reflects some sample behavior. Thus, it is necessary to consider its incompleteness or lack of representativeness (van der Aalst, Adriansyah, de Medeiros, et al., 2012).

### 4.2. DATA COLLECTION

The data for this study was extracted from a well-known LMS, Moodle, the learning platform adopted by NOVA IMS. The download and preparation of the event data from the Moodle System occurred in July 2021. The resulting event log is in the form of a CSV file, with a sheet referent to the event data from 2020 and another for 2021 until the extraction date. In other words, from the year 2020, we have the exam season of the 2019/2020 1st semester, the entire 2nd semester of 2019/2020 and then the 1st semester of 2020/2021 but without the exam season. From the year 2021, we have the exam season of the 1st semester of 2020/2021 and the entire 2nd semester of 2020/2021, which corresponds to three whole semesters.

The event log consists of seven attributes, a description of each attribute can be seen in the succeeding Table 2.

Table 2 - Description of attributes from Moodle event log

| Attribute          | Description                       |
|--------------------|-----------------------------------|
| <b>timecreated</b> | Date and time of the event        |
| <b>eventname</b>   | Name of the activity conducted    |
| <b>action</b>      | Activity performed by the student |
| <b>target</b>      | Name of the event target          |
| <b>userid</b>      | The user id of the student        |
| <b>courseid</b>    | Assigned id of each course        |
| <b>coursename</b>  | Name of the course                |

The three main attributes necessary for the implementation of process mining are the timestamp, case ID and activity. In this case, these are *timecreated* (format: “dd/mm/yyyy hh:mm”), *userid* and *eventname*, correspondingly. With the inspection and familiarization of the event log, it is possible to recognize that, for this study, the *coursename* attribute is similarly essential in order to distinguish the

events of each course. On the other hand, it is already possible to delete some attributes, thus starting the first step of preprocessing, choosing and filtering the necessary attributes. Therefore, we have three attributes that can be excluded: *action*, *target* and *courseid*, since the first two are included in the *eventname*. The following preprocessing steps include corrections and adjustments, such as correcting errors in data for the *coursename* attribute, deleting courses belonging to previous years and dividing the event log into semesters and subsequently by courses.

Before proceeding to the data processing stage, a careful analysis of the data is necessary. The data analysis includes information about the activities and their events, such as the total number of events, their occurrences in different cases and courses, the number of students per class, the various activities present in different scenarios, or more specifically, which courses have the most events. This analysis of the statistics provides insights into the size of the process and the event. Moreover, it helps to adjust mining algorithms and evaluate the results obtained in the coming phases (Bozkaya et al., 2009).

### **4.3. DATA PROCESSING**

The data were explored, preprocessed and analyzed. Then, it follows the decision and selection phase in order to choose accurate, representative and significant event data for the process mining application. In light of this, we start by verifying which courses have the most events, which can be seen in Appendix I. The 20 courses with the most events in 2020 and 2021 are displayed in the form of two packed bubble visualizations, respectively.

#### **4.3.1. Data Selection**

The exploration of courses with data available in the two academic years, 2019/2020 and 2020/2021, led the research to pursue a comparative analysis between one semester in 2019/2020 and one in 2020/2021, specifically the 2nd semester. Unfortunately, due to missing data or data errors, only three courses are available to be analyzed according to these criteria, which can be noticed in Appendix I. In addition, a significant highlight of the registration with the name “NOVA IMS Moodle” can be detected, which we assume to have been a system error. Further research showed that most of these registrations were in the 1st semester of the year 2020/2021 and had no further details, which also explains the absence of subjects of the 1st semester in the graph.

As a result of this comparison, the first step in the data processing stage emerged, the manual clustering, which focuses on identifying and grouping students who repeated the course. This was achieved by identifying students who participated in at least one quiz in the first year and joined again in the second year; thus, we will assume it corresponds to students who failed the course. By searching the courses with the highest number of students who repeated the course in the second year, it was concluded that only one of the three courses had a significant number of failing students, Econometrics, which belongs to the 2nd year and 2nd semester of the Information Management Degree. Therefore, we chose to analyze the Econometrics course since it has a considerable number of failing students and a larger sample.

#### **4.3.2. Econometrics**

Analyzing only the logs from the 2019/2020 academic year, the Econometrics course has 98 students. Using the manual clustering approach, it is possible to detect two groups from the 98 students: 11

students who participated in the quizzes in the two academic years (failing students) and 87 students who only participated in the quizzes in the 2019/2020 academic year (passing students).

At this stage, additional filtering was executed, where activities considered irrelevant to the students' performance were excluded. Thus, of the activities saved by Moodle (20 in total), only 13 activities were used (see Table 3). With this filter, the log file is reduced from 38 479 events to 37 995 events.

Table 3 - Description of the activities considered relevant to the students' performance

| Activities                         | Description                             |
|------------------------------------|-----------------------------------------|
| <b>course viewed</b>               | Course opened                           |
| <b>folder all files downloaded</b> | All files downloaded                    |
| <b>folder viewed</b>               | Folder opened                           |
| <b>forum viewed</b>                | Forum opened                            |
| <b>forum discussion viewed</b>     | Forum discussion opened                 |
| <b>quiz attempt reviewed</b>       | Review results and feedback of the quiz |
| <b>quiz attempt started</b>        | Quiz started                            |
| <b>quiz attempt submitted</b>      | Quiz result saved                       |
| <b>quiz attempt viewed</b>         | Questions in the quiz are displayed     |
| <b>quiz course module viewed</b>   | Quiz front page opened                  |
| <b>quiz summary viewed</b>         | Quiz summary before saving is displayed |
| <b>resource viewed</b>             | Resource opened                         |
| <b>url viewed</b>                  | URL opened                              |

In addition to the manual clustering, we propose using automatic clustering as a processing task involving grouping students based on their interactions with the Moodle course. The suggested approach was inspired by the approach developed by Romero, Cerezo, Bogarín and Sánchez-Santillán (2016). The clustering application aims to group students with similar results or characteristics in favor of the process mining activity to generate more specific models depending on students' behavior. Before going into more detail about the application of automatic clustering, it is necessary to acknowledge that this technique was implemented in an already reduced event log with selected variables and filtered activities.

The automatic clustering uses the Moodle usage data through the reports of each student's interaction in Moodle. It is based on variables related to the time spent studying, the time spent in activities related to the quiz, the number of events in various scenarios, the number of activities and the number of days that Moodle was accessed (more details in Table 4).

The variables used in the clustering algorithm were calculated through Excel, Celonis (Celonis, 2020) and others directly from the statistics automatically originated by the process mining tool, Disco (Fluxicon, 2021), as displayed in Figure 9 below.

For instance, the variables related to time were calculated using Excel. For example, the variable *time theory*, which describes the time spent studying, is calculated by subtracting the time the student accessed the activities *resource view* or *folder view* and the next different activity, followed by adding these periods together. In addition, concerning this variable, it is necessary to consider that it is not entirely reliable, as students could be doing other actions simultaneously or on other websites. Furthermore, the time spent studying is a relative variable, considering it is not necessarily linked to the results obtained; it depends on the quality of the study. Hence, the importance of the other variables included in the analysis and their correlation.

Concerning the variables derived from the Celonis tool, unlike the Disco tool, it is necessary to build an overview board with the desired statistics, where it is possible to add various components, from charts and tables to single KPIs and selection components. Therefore, to calculate the variables that were considered necessary, we built the overview board that can be accessed in Appendix II. For example, the variables *relevant events*, *theory events* and *quiz events* were calculated by selecting the corresponding activity names and taking the values obtained from the table that displays the number of events by each *userid*. This can also be seen in Appendix II, regarding the variable *quiz events*, where it is possible to visualize the change of values and tables due to the selection. As for the *days in Moodle* variable, it was necessary to create the formula: `COUNT(DISTINCT ROUND_DAY("201920_S2_-_Econometria_I_csv"."TIMECREATED"))` and which can then be accessed in the table with the *userid* or as a number single KPI.

Table 4 - Description of the variables used for clustering

| Variable name              | Description                                                            | Extraction or Calculation Method                                                                                  |
|----------------------------|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| <b>Relevant events</b>     | Number of events within the relevant activities                        | Total of relevant events (Disco and Celonis)                                                                      |
| <b>Relevant activities</b> | Number of relevant activities performed                                | Total of relevant activities (Celonis)                                                                            |
| <b>Days in Moodle</b>      | Number of days Moodle was visited (in days)                            | Total number of days Moodle was visited (Celonis)                                                                 |
| <b>Activity days</b>       | Number of days between the first and the last relevant event (in days) | Difference of the date of last relevant activity and date of first relevant activity (Disco and Celonis)          |
| <b>Time theory</b>         | Total time in minutes spent on theoretical activities (in minutes)     | Sum of the periods between <i>resource viewed</i> or <i>folder viewed</i> and the next different activity (Excel) |
| <b>Theory events</b>       | Number of events spent on theoretical activities                       | Total of theoretical activities considered (Celonis)                                                              |
| <b>Time quiz</b>           | Total time spent on quiz activities (in minutes)                       | Sum of the periods between quiz activities and the next different activity (Excel)                                |
| <b>Quiz attempts</b>       | Number of quiz attempts                                                | Total of quiz attempts (Celonis)                                                                                  |
| <b>Quiz events</b>         | Number of events spent on quiz activities                              | Total of events spent on quiz activities (Celonis)                                                                |

The next phase involves converting this information stored in a CSV file into an attribute-relation file format (ARFF), the format required by Weka, the open-source data mining tool used to elaborate the automatic clustering. Weka comprises a collection of machine learning algorithms for data mining tasks, such as data preparation, classification, regression, clustering and visualization (Witten, Frank, Hall, & Pal, 2016). The choice of this machine learning workbench written in Java is due to the tool's capability of processing data obtained in the structure of a single relational table and from it obtain valuable information and generate an accurate predictive model (Frank, Hall, Trigg, Holmes, & Witten, 2004). Additionally, WEKA is widely recognized by academia and business (Hall et al., 2009).

Regarding the clustering, the Expectation-Maximization (EM) clustering algorithm was chosen in this study. EM algorithm aims to maximize the likelihood of the data (RapidMiner, 2021). EM is similar to the K-Means technique; however, it does not require specifying the number of clusters (Cristóbal Romero et al., 2016). When using this clustering algorithm, the objective is to group students together that have similar characteristics and behavior when using Moodle.

Clustering based on the variables described above resulted in three clusters, distributed as follows:

- Cluster 0: 46 students (2 failing students)
- Cluster 1: 44 students (3 failing students)
- Cluster 2: 8 students (6 failing students)

Table 5 reports the statistics, the mean and standard deviation, for each of the nine attributes in the three detected clusters provided by the output window of Weka.

Table 5 - Statistics (means  $\pm$  std. dev.) of each of the three clusters

| Attribute                  | Cluster 0           | Cluster 1             | Cluster 2           |
|----------------------------|---------------------|-----------------------|---------------------|
| <b>Relevant events</b>     | 306.95 $\pm$ 64.33  | 526.5 $\pm$ 144.49    | 109.45 $\pm$ 92.61  |
| <b>Relevant activities</b> | 11.6 $\pm$ 1.06     | 12.05 $\pm$ 1.1       | 8.51 $\pm$ 3.43     |
| <b>Days in Moodle</b>      | 30.998 $\pm$ 9.46   | 48.08 $\pm$ 17.02     | 9.41 $\pm$ 6.01     |
| <b>Activity days</b>       | 127.35 $\pm$ 11.1   | 135.97 $\pm$ 8.48     | 109.03 $\pm$ 25.97  |
| <b>Time theory</b>         | 772.99 $\pm$ 509.28 | 1773.14 $\pm$ 1197.45 | 108.77 $\pm$ 132.44 |
| <b>Theory events</b>       | 46.08 $\pm$ 17.47   | 81.01 $\pm$ 31.23     | 15.04 $\pm$ 9.25    |
| <b>Time quiz</b>           | 186.24 $\pm$ 46.83  | 335.71 $\pm$ 106.57   | 47.09 $\pm$ 42.01   |
| <b>Quiz attempts</b>       | 2.58 $\pm$ 0.57     | 3.40 $\pm$ 0.58       | 0.75 $\pm$ 0.66     |
| <b>Quiz events</b>         | 153.6 $\pm$ 49.39   | 239.7 $\pm$ 99.11     | 65.34 $\pm$ 87.78   |

In the interest of gaining insights on the distribution of students' activities in Moodle, throughout the semester before executing the process mining application, two calendar heatmaps were developed in Excel (Figure 8). The visualization on the left represents the total number of activities performed in Moodle on each day of the semester. Additionally, the horizontal and vertical bars represent the activities on each week and each day of the week, respectively. For example, the four days where

quizzes were held stand out, and through the vertical bars below it is possible to notice less activity on the weekends.

On the right side of Figure 8, we have the same representation but without the quiz days, which are marked in white. Therefore, it is possible to notice the activity on the remaining days, such as a more significant presence in the days immediately before the quiz, especially in April and June, and a greater frequency of visits on Mondays and Thursdays.

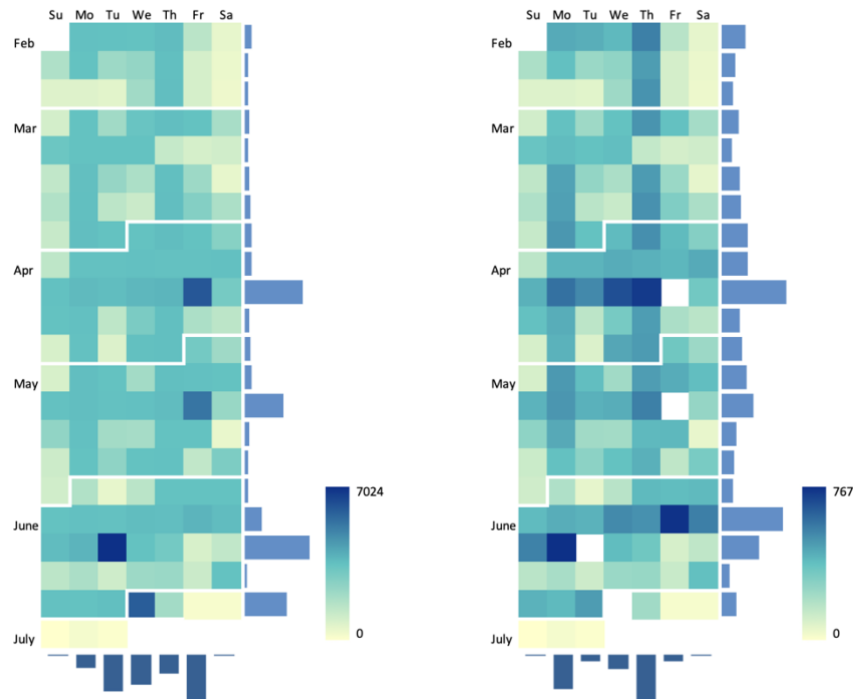


Figure 8 - Calendar heatmaps of students' activity

#### 4.4. PROCESS MINING AND ANALYSIS

As previously mentioned, process mining is a fundamental tool to understand the underlying process, its performance and relevant patterns. In this study, we intend to use the benefits of process mining to discover patterns in students' behavior when using Moodle, the LMS used by NOVA IMS. We aim to overcome the barrier originated by web-based education and understand how students use Moodle to learn and how they interact with the tool and the course materials.

Consequently, to generate the process models that will allow a better comprehension of the learning process and its results, we use the Disco tool, which is based on the Fuzzy Miner algorithm (Fluxicon, 2020). The tool was chosen because it incorporates an automated discovery process that can handle complex and heterogeneous event logs. Moreover, it has a wide variety of filters and metrics available, making it easier and more iterative to explore the event log and the process. Furthermore, the tool offers intuitive and modern visualizations with factual and accurate representation. Finally, Disco allows to straightforwardly adjust the level of detail appearing in the process map with two sliders, activities slider and paths slider (Fluxicon, 2021).

This process discovery phase began with the attribution of the Disco requirements for an event log: Case ID, Activity and Timestamp, which in our event log corresponds to *userid*, *eventname* and *timecreated*. Then, the algorithm was applied over each of the following logs:

- All students (98 students)
- Students who passed (87 students)
- Students who failed (11 students)
- Students assigned to Cluster 0 (44 passed and 2 failed)
- Students assigned to Cluster 1 (41 passed and 3 failed)
- Students assigned to Cluster 2 (2 passed and 6 failed)

First, the process models were generated. Next, the models discovered using these logs were described and analyzed.

The first considerations to notice are from the screenshot of the Statistics view in Disco when using the log file with all students. By observing Figure 9, it is possible to acknowledge that the number of cases is precisely the same as the number of variants; in other words, each of the 98 students has its own unique path. It is also possible to obtain other information about the event log, such as the number of events, the beginning and end dates, which correspond to the beginning and end of the semester. Finally, as in the calendar heatmap, we can discern the four peaks on the events over time, corresponding to the quiz days.

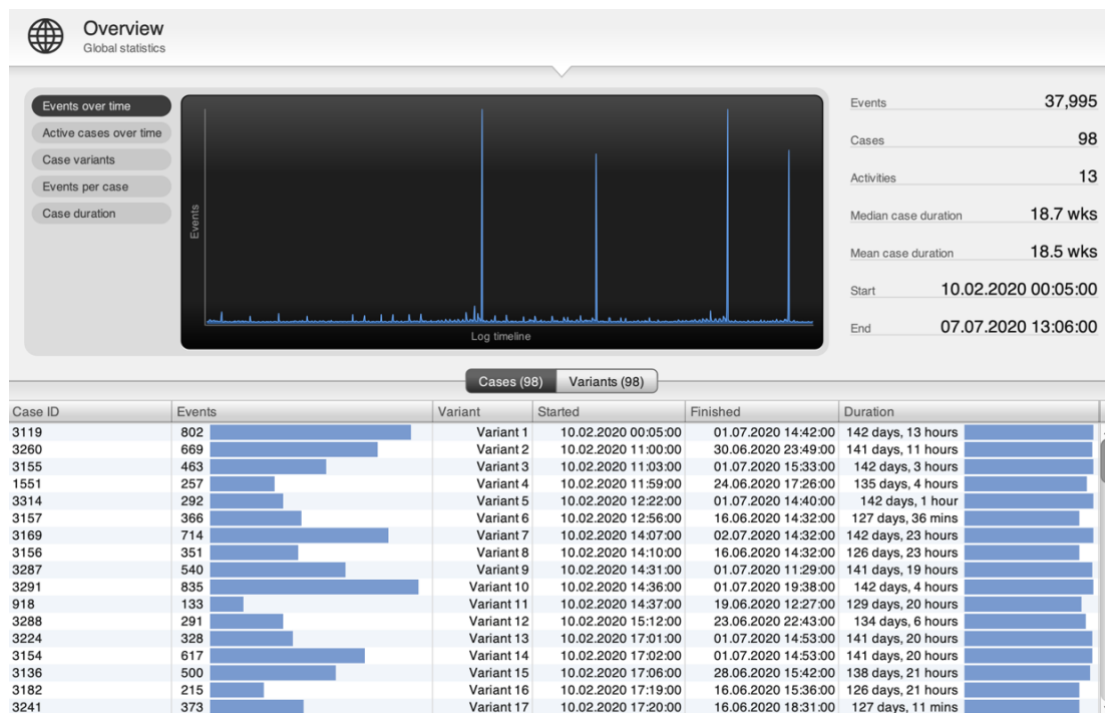


Figure 9 - Disco's Statistics view

The discovery process, including events from all students, resulted in the spaghetti process map in Figure 10. Spaghetti processes is the term used to describe unstructured and complex processes with many activities and even more arcs, making it challenging to observe patterns and find insights (van der Aalst, 2016). The event log had already been filtered to contain only the most relevant activities,

and the view sliders were fixed to show all activities that occurred (100%) and 90% of the connections between the activities; nevertheless, the process remained excessively complex and difficult to interpret.

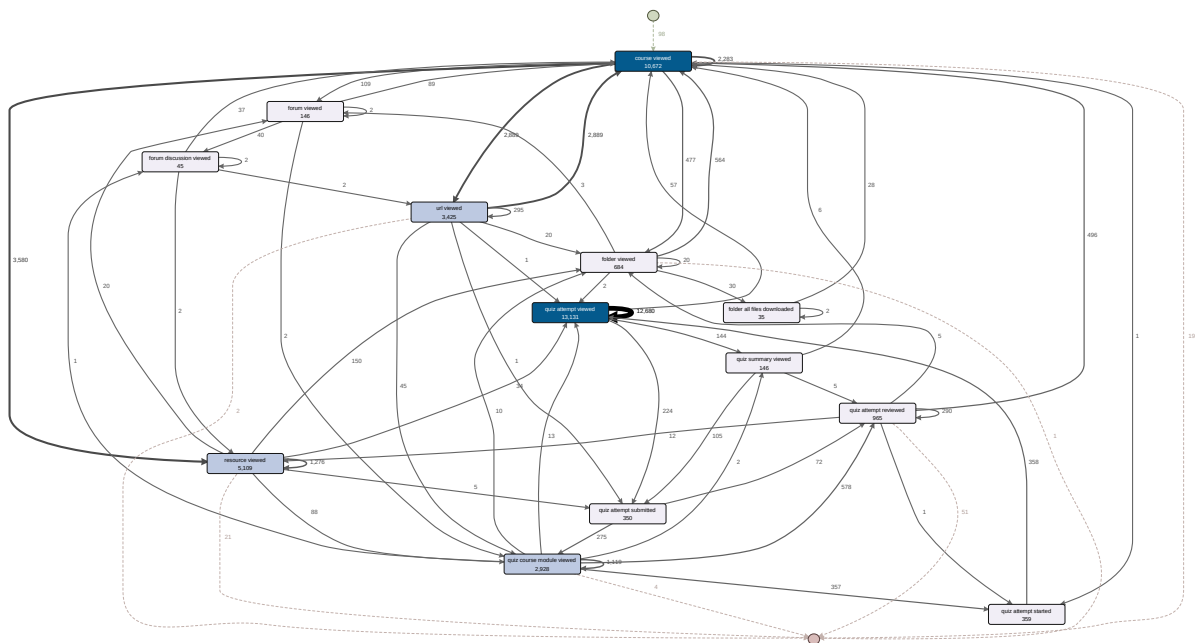


Figure 10 - Process model of all students with 100% activities and 90% paths

Therefore, to simplify the process models, but simultaneously have the same level of complexity in all the consecutive process maps for a fairer comparison, we fixed the sliders to show all activities that occurred (100%) and 40% of the connections between the activities. After experimenting with different percentages and levels of granularity, we realized that the best results came from process maps with all activities and only the main or necessary flows. Thus, the model includes paths with less frequency as long as they are the essential paths to keep all activities and their connections.

Contemplating again, in Figure 11, the process model referring to all students, but this time with 40% of connections and displayed with the case frequency view instead of the default view, which is the absolute frequency view. The absolute frequency is based on total frequencies, that is, the total number of times an activity or path was performed or followed. In contrast, the case frequency shows relative numbers without counting the number of repetitions within a single case (Fluxicon, 2020). Accordingly, in Figure 11, the numbers represent the number of students who executed each activity or followed each path rather than the number of times the activity or connection was performed in total, enabling an initial overview and perception of the students' participation.

We can see that all 98 students accessed the course page in Moodle and viewed the courses' resources, but three of them did not participate in the quizzes. Only 20 students downloaded all the files, and just over half accessed the course forum page. Furthermore, it is detectable that relatively often, the students did not submit the quiz and let it be submitted automatically when the time limit is over. This can be identified by observing the path that goes directly from *quiz attempt viewed* to *quiz attempt submitted* without going through the *quiz summary viewed* page. From Figure 11, we can see that 89 students did this at least once.



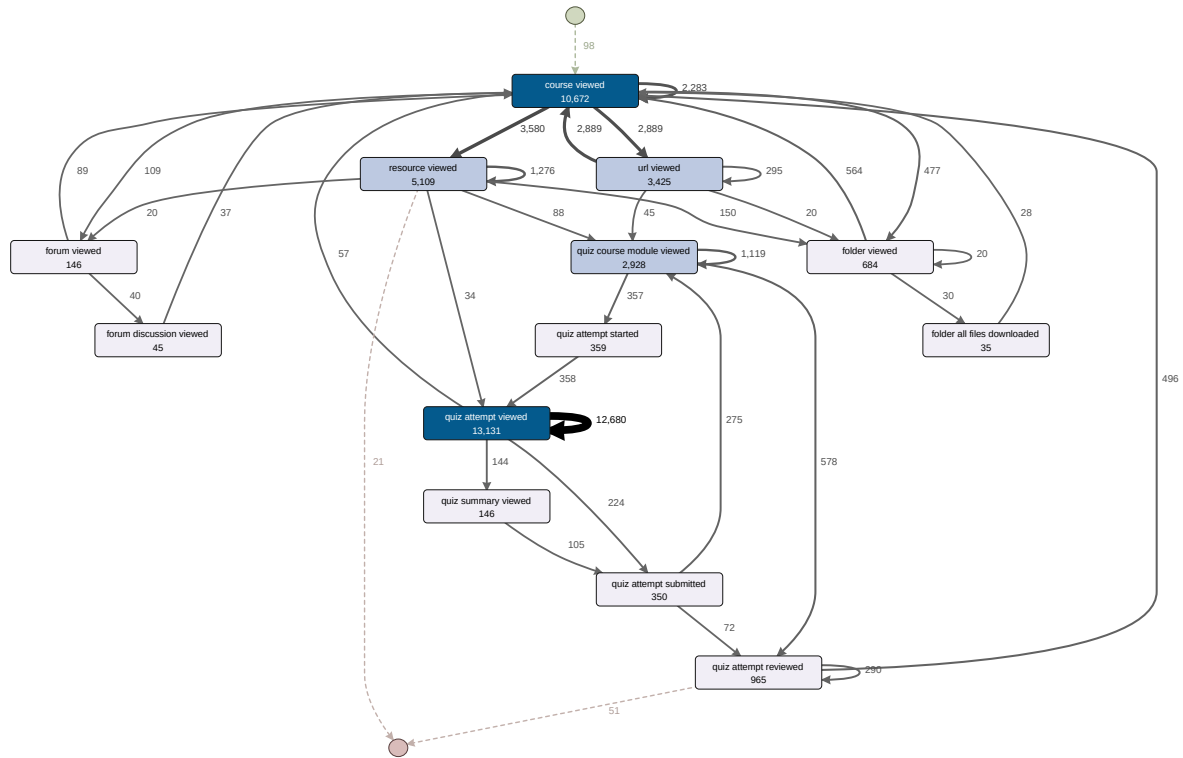


Figure 12 - Process model of all students with 100% activities and 40% paths

The next step in our analysis approach targets the other subsets of logs already mentioned, which, due to the filtering and the clustering performed, are better controlled processes, leading to more specific process models. Beginning by comparing the process map seen above (Figure 12) with the map of the students who passed (Figure 13), it is possible to recognize that the similarities are evident, the only meaningful difference being the number of events of each activity and path.

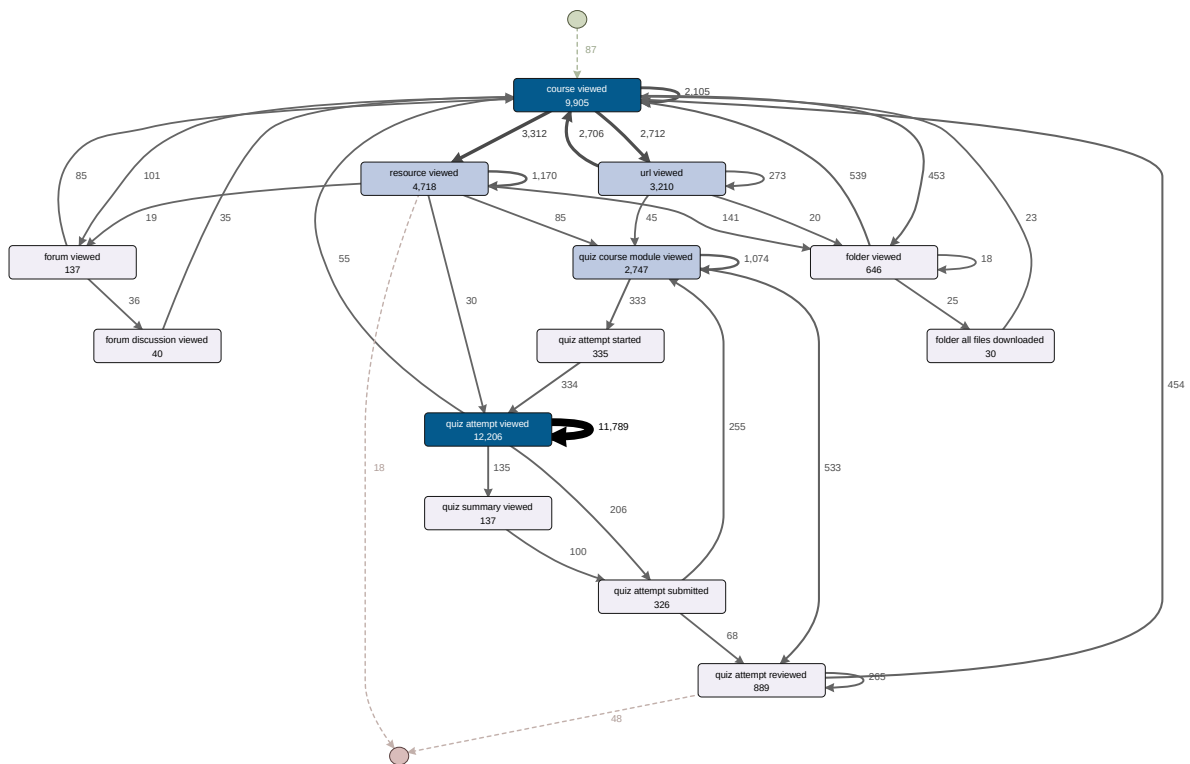


Figure 13 - Process model of passing students with 100% activities and 40% paths

It follows the process maps of the failing students, which is a simpler model than the ones analyzed so far. On account of two motives, it contemplates only 11 cases, and most of these 11 students interacted less and performed fewer activities in Moodle. This model detected the lack of links between activities, making the activity groups stand out, such as quiz, folder, and forum activities.

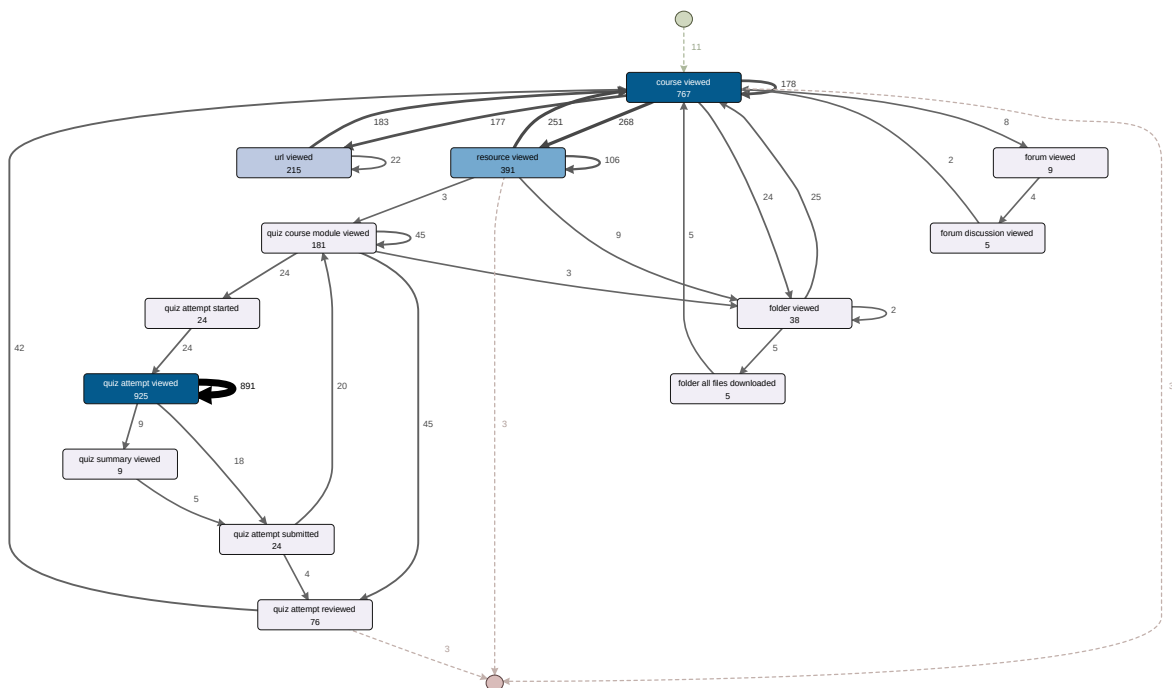


Figure 14 - Process model of failing students with 100% activities and 40% paths

Apart from that, the Disco tool can also display more specific characteristics of each activity and path. For example, when clicking on an activity in the process map, it displays all metrics in an overview badge. Comparing and analyzing the activity *quiz summary viewed* metrics between the previous two groups, passing and failing students, we can see a significant difference in the percentage of case coverage, which shows that in the first case 64 of the 87 students belonging to the group displayed the quiz summary before submitting it. While in the second group only three of the 11 students performed this activity.

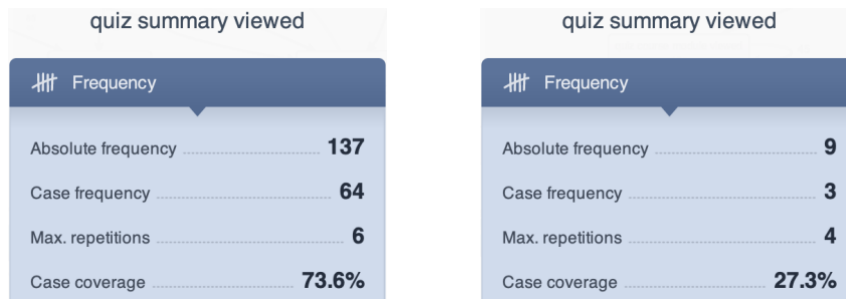


Figure 15 - Overview badge of quiz summary viewed activity from passing students (left) and failing students (right) models

It follows the analysis of the three process models originated through the logs of each cluster aggregated at different levels of abstraction based on the nine variables above described. In a general observation of the three process maps, the different behavior patterns pursued by the students in each cluster are perceivable.

We will now contemplate the two process models of Clusters 0 and 1, which are constituted by a majority of successful students, with approximately 96% and 93% of approved students, correspondingly. Regarding the distribution of students in each of the clusters, we have that Cluster 0 has two more students than Cluster 1. However, we see in Table 5 that Cluster 1 presents higher values in the mean and standard deviation of all variables, particularly in the number of events and in the number of days students visit Moodle, with the first taking an average of only 31 days and the second 48 days. This discrepancy in the access to Moodle is also noticeable in the process maps of each cluster, specifically in the number of times each activity was executed, which is significantly higher in every activity of the Cluster 1 process.

Furthermore, we can notice in the process model of Cluster 0 that it has many connections; however, most of them are small paths that mainly connect activities close to one another and belong to the same activity group. While in the process of Cluster 1, it is possible to notice more extensive connections and a more frequent interconnection between activities of different groups, only activities related to visiting the forum page stand out from the rest.

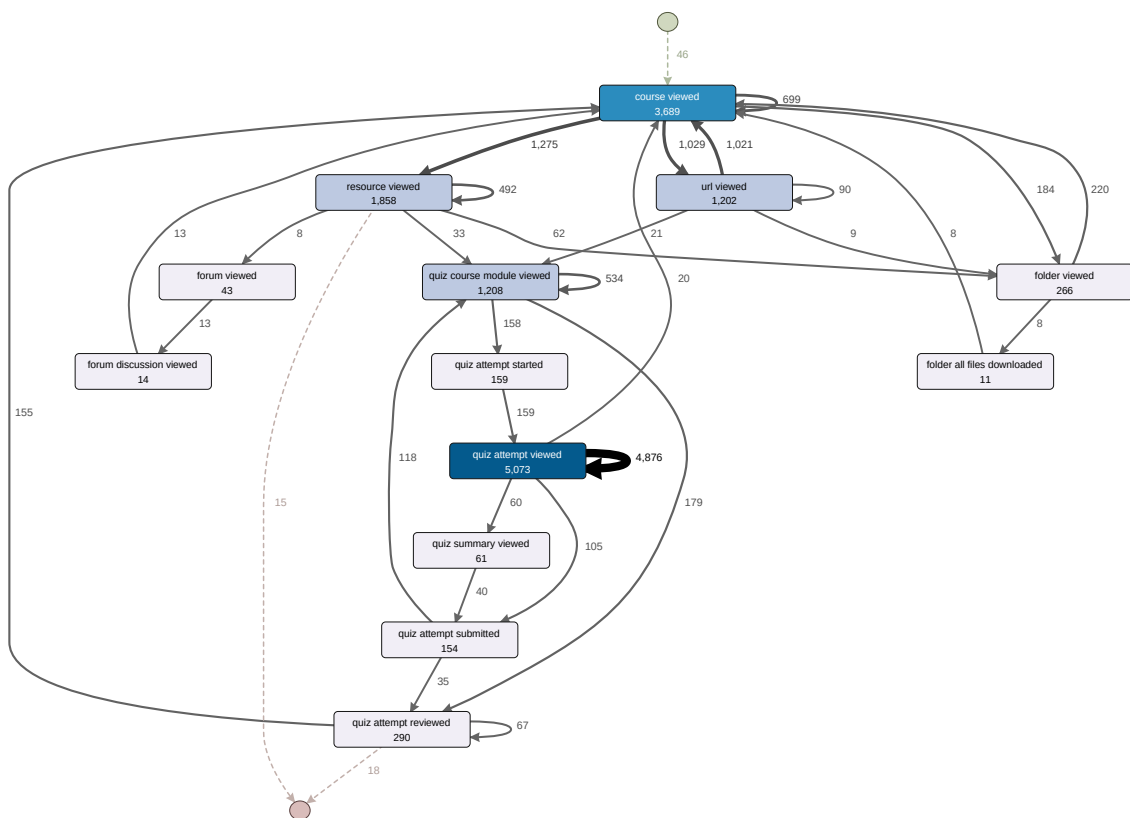


Figure 16 - Process model of Cluster 0 with 100% activities and 40% paths

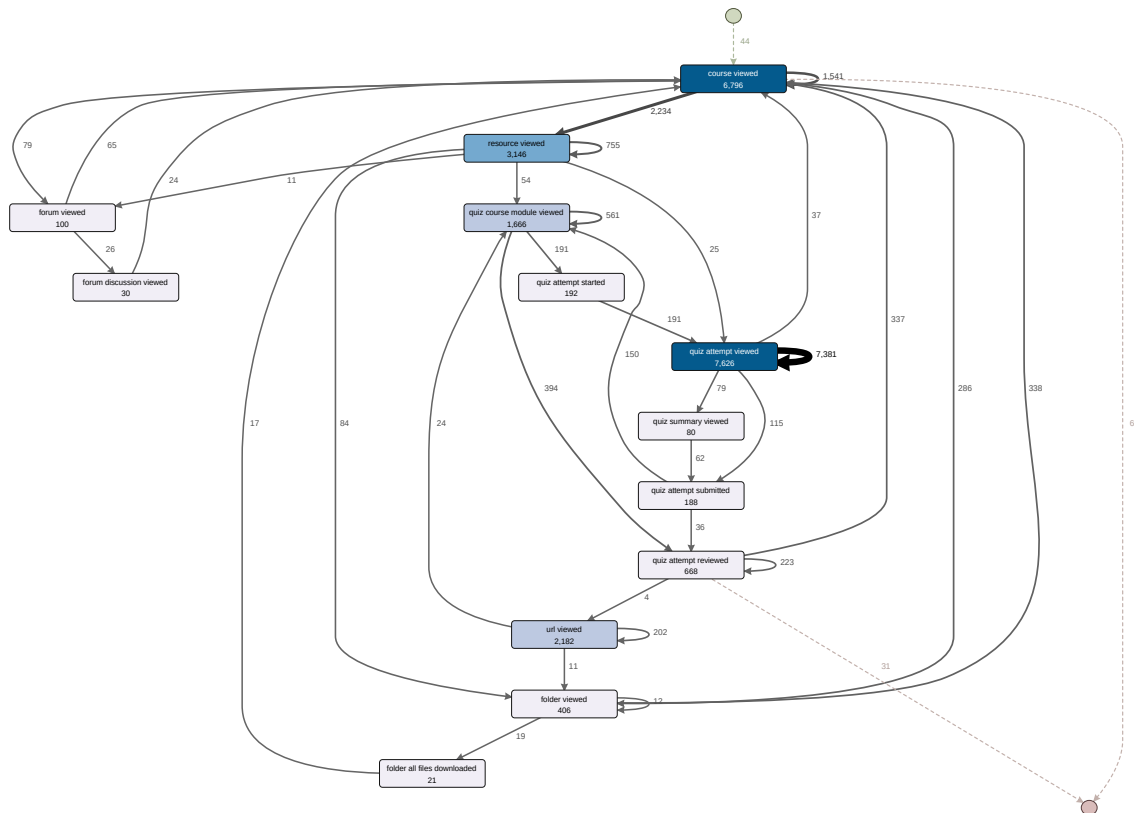


Figure 17 - Process model of Cluster 1 with 100% activities and 40% paths

The process of Cluster 2, visible in Figure 18, shows a significant contrast regarding the last two models, being a considerably simpler process. The group of students contains six of the 11 students who failed; therefore, Cluster 2 is constituted by 75% of failing students. In this model, we see that the activity *course viewed*, despite continuing to be a central activity, does not stand out as much, being, in this case, the main activity the *quiz attempt viewed*. In addition, we can observe that the activities are concentrated into two groups. The first corresponds to the group of activities related to the quiz, and the second includes all the rest, expressing a greater focus and emphasis on quiz-related activities.

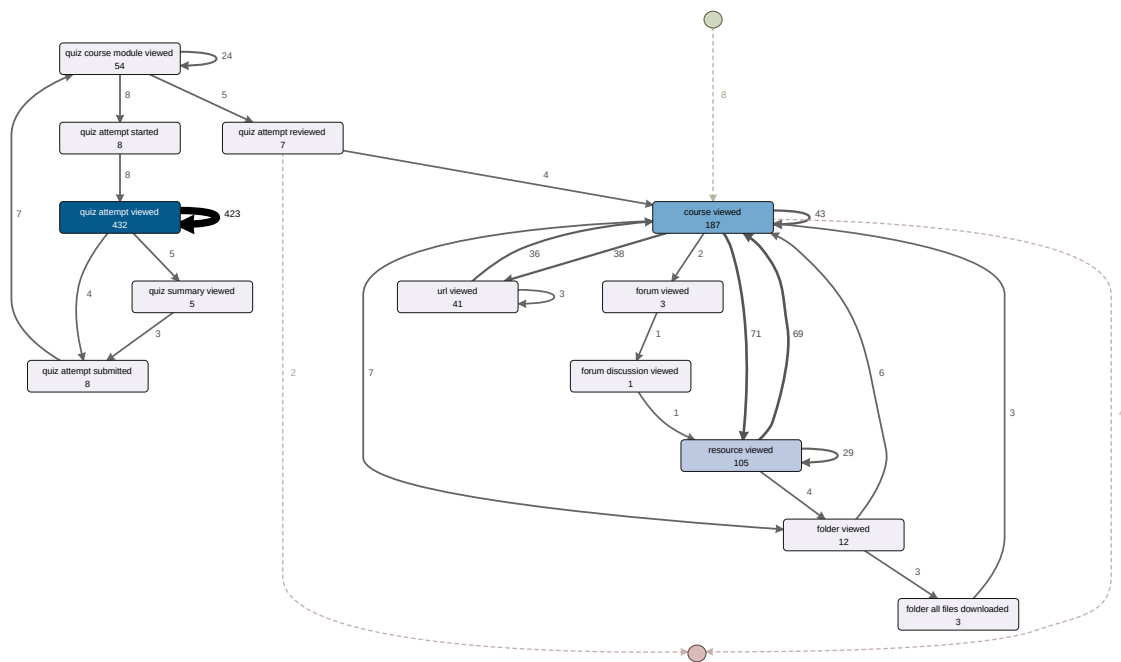


Figure 18 - Process model of Cluster 2 with 100% activities and 40% paths

Finally, it is essential to analyze this last model with the process map that only includes failing students, comparing Figures 14 and 18. The process model of failing students presents more events and denotes more action from these students than the process model of Cluster 2, as expected, considering it includes more students and the ones included in Cluster 2 were grouped based on the activity and interaction in Moodle. Nevertheless, there are similarities between the models, from being simple processes enabling a straightforward analysis to the prominence of the quiz activities and a strong interconnection between the remaining activities.

#### 4.5. EVALUATION

This section tests the previously used approach on the dataset covering the second academic year accessible, 2020/2021, and relating to the same course, Econometrics. We intend to verify the possibility of predicting students at risk of failing and evaluate its usefulness through the combination of automatic clustering and process mining, in order to analyze and recognize potential predictable patterns. A viable application can contribute to early actions resulting from the premature detection of difficulties and prediction of students at risk of failing.

In the 2020/2021 academic year, the Econometrics course had 80 students, of which only 69 participated in the quizzes, as evident in Figure 19, and our study will focus only on the latter. As before, the analysis will focus on the main activities of the process, which are summed up to 12 activities. Accordingly, the event log without the 11 students who did not participate in the quiz and only containing the main activities constitutes 22 649 of the 24 288 events.

Regarding clustering, the strategy for grouping students differs from the previously established one, since it is impossible to identify the students who failed in the present case; therefore, only automatic clustering will be performed. The automatic clustering was performed based on the usage and interaction patterns of students on the Moodle platform comprehended in the variables formerly detailed.

We used the EM clustering algorithm based on the nine variables, similar to what was previously executed. This was followed by an iterative process aiming to obtain clusters and, consequently, models that fit the data as much as is viable and pursued through the option available in the Weka tool “ignore attributes”. The measurement applied used the log-likelihood values presented and automatically calculated by Weka. The log-likelihood is a method for probabilistic clustering methods that measures the log-likelihood of the obtained clusters on the training data: the higher this value, the better the model fits the data (Witten, Frank, Hall, & Pal, 2016). This resulted in the following distribution, in four clusters, based on seven of the nine variables, since the variables *relevant activities* and *activity days* were ignored:

- Cluster 0: 26 students
- Cluster 1: 9 students
- Cluster 2: 30 students
- Cluster 3: 4 students

The mean and standard deviation for each of the seven attributes in the four detected clusters provided by the output window of Weka are given in Table 6.

Table 6 - Statistics (means  $\pm$  std. dev.) of each of the four clusters

| Attribute              | Cluster 0           | Cluster 1             | Cluster 2          | Cluster 3           |
|------------------------|---------------------|-----------------------|--------------------|---------------------|
| <b>Relevant events</b> | 369.48 $\pm$ 68.23  | 511.8 $\pm$ 60.23     | 275.65 $\pm$ 46.11 | 132 $\pm$ 29.95     |
| <b>Days in Moodle</b>  | 36.65 $\pm$ 12.22   | 60.99 $\pm$ 10.55     | 38.7 $\pm$ 8.3     | 17.5 $\pm$ 7.89     |
| <b>Time theory</b>     | 938.23 $\pm$ 877.39 | 4307.75 $\pm$ 3349.76 | 842.6 $\pm$ 703.97 | 170.75 $\pm$ 152.95 |
| <b>Theory events</b>   | 45.82 $\pm$ 16.41   | 112.17 $\pm$ 22.59    | 45.14 $\pm$ 21.44  | 22.5 $\pm$ 7.7      |
| <b>Time quiz</b>       | 242.94 $\pm$ 54.25  | 238.97 $\pm$ 79.37    | 166.55 $\pm$ 34.71 | 63 $\pm$ 13.36      |
| <b>Quiz attempts</b>   | 2.83 $\pm$ 0.38     | 2.64 $\pm$ 0.48       | 2 $\pm$ 0.0002     | 1 $\pm$ 0.59        |
| <b>Quiz events</b>     | 185.83 $\pm$ 56.21  | 188.25 $\pm$ 46.69    | 109.66 $\pm$ 18.4  | 42.25 $\pm$ 4.97    |

It follows, once again, the development of process models through the Disco tool and the corresponding analysis of the generated models. The sliders are also fixed to show 100% of the activities and 40% of the paths.

In contemplation of an overview of the event log under analysis, we begin by inspecting the simplified process map of all students, covering the 12 main activities, with 94% of the events, and presented in the case frequency view. As noted above, it can be identified from Figure 19 that 11 of the 80 students did not participate in any of the quizzes performed; on the other hand, every student accessed the course's resource page. In addition, the low turnout to the course forum is notorious, only 28 students accessed it, and even less frequent was the download all files function, being performed by only 16 students.

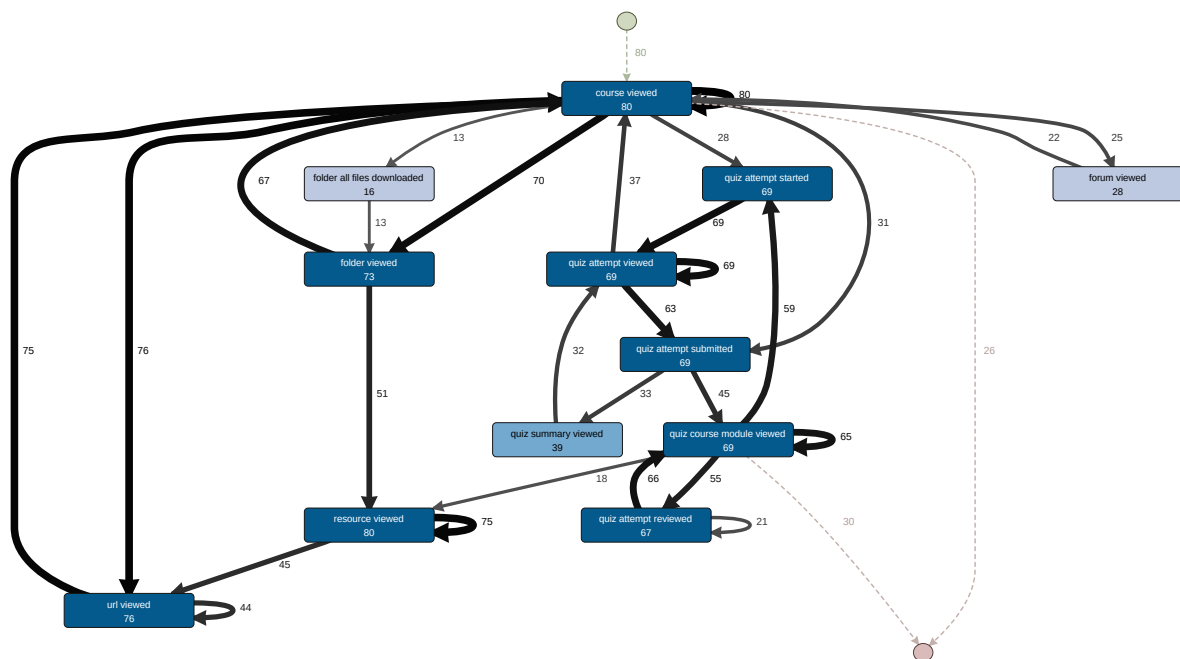


Figure 19 - Process model of all students, academic year 2020/2021 (case frequency view)

With regard to the process models related to the generated clusters, we will address the last three clusters in virtue of a comprehensive but succinct analysis, and this selection was based on the statistics presented in Table 6. We will include, therefore, the group with more activity and interactions (Cluster 1), the one that presents the lower values in all variables (Cluster 3) and the cluster that, due to the indicators close to those of the previous group, we consider critical to include and analyze the model (Cluster 2).

The process of Cluster 1 can be found in Figure 20. Despite representing the activity of only nine students, it is a compound map, with many connections between each activity and with high frequency values. The activities related to entering the main page course and performing the quiz (*quiz attempt viewed* activity) stand out. There is also a third activity with significant values, corresponding to visiting the resources page action. From the observation of paths, it is possible to notice that it happened four times that the resources page was visited immediately before or during the quiz.



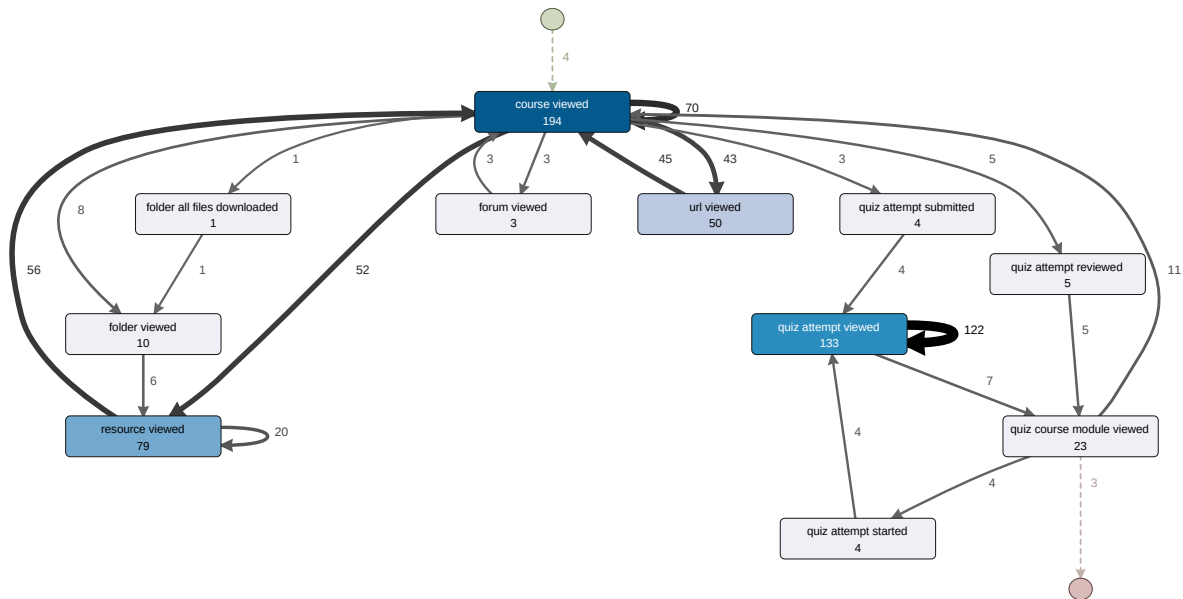


Figure 21 - Process model of Cluster 3, academic year 2020/2021

Cluster 2 comprises 30 students, a significant part of the students in the course. Nevertheless, comparing Figure 22 that illustrates the process model of Cluster 2 with Figure 21, the similarity is evident, from a process model considered simple to the clear division of two activities groups, quiz and the rest. The latter was a characteristic also noted in the process models of the 2019/2020 academic year, more precisely in the failing students and Cluster 2 models. In addition to the typical subprocess related to the quiz in all models, in the model illustrated in Figure 22, it is possible to notice a well-defined subprocess related to the consultation of materials in the course, indicating the existence of a more common path in material research and in the study process.

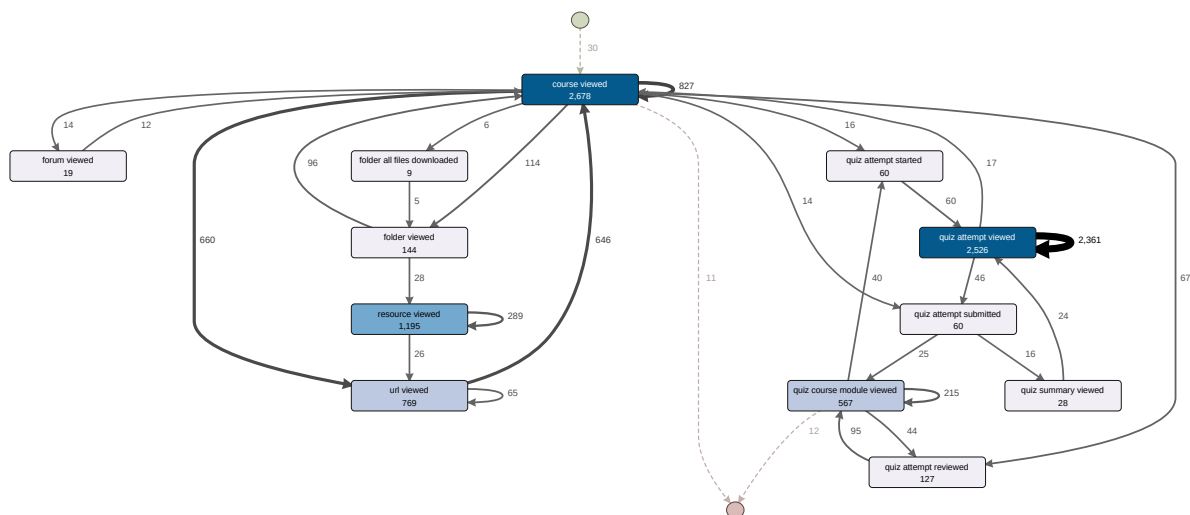


Figure 22 - Process model of Cluster 2, academic year 2020/2021

In addition to this prediction, it is also possible to analyze the behavior in Moodle of students who were repeating this year (Figure 23), that is, students whose activity in Moodle in the 2019/2020 academic year is represented in Figure 14. The comparison between the two figures shows a lesser

emphasis on the activities related to the quiz and greater interconnection between the various activities on this second process. Furthermore, through the *days in Moodle* variable, available in the Celonis tool, we calculate that, in the year 2020/2021, these students entered Moodle 10 days more than in the previous academic year, from 97 to 107 days. Also, the number of events was higher this second year, from 2669 to 2826 events.

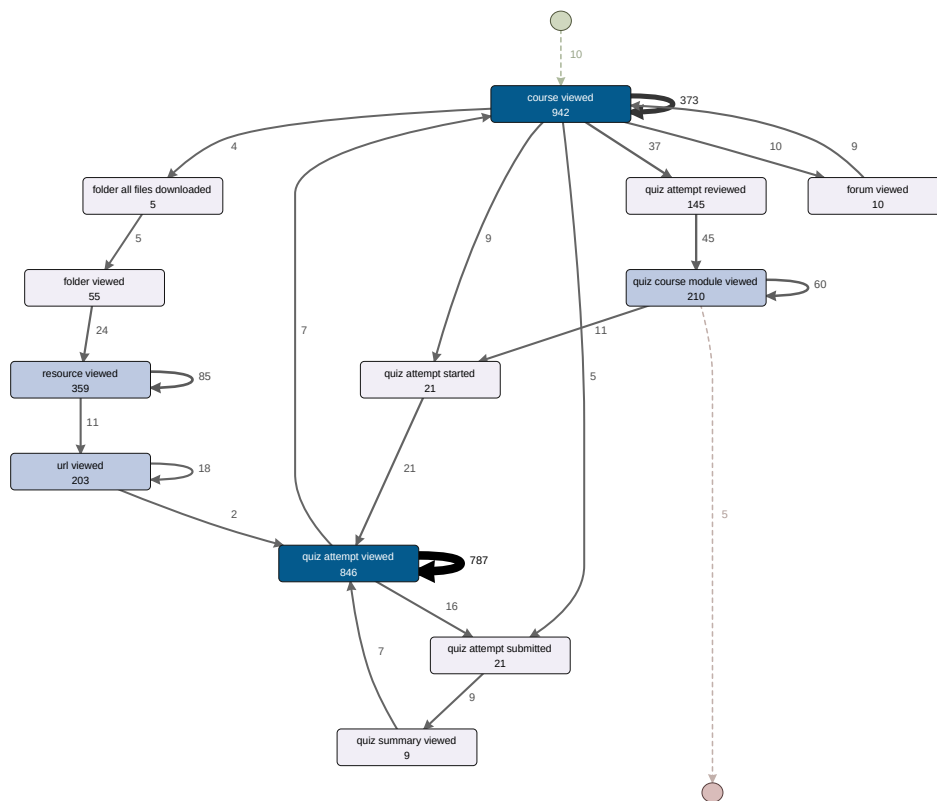


Figure 23 - Process model of repeating students, academic year 2020/2021

## 4.6. DISCUSSION

#### 4.6.1. Evidence and Outcomes

This study proposed the use of process mining techniques in conjunction with clustering to analyze students' behavior in the LMS Moodle. Different models were obtained by grouping the data logs according to students' success and interactions in Moodle.

One of the first observations we obtained was that the educational process under study does not have a most frequent path. The variants are different for each case, demonstrating the diversity and complexity of the process and students' diverse behavior. This initial consideration ended up defining the rest of the process exploration.

In the model that involves all students, it was observed that the vast majority of the students let the quiz be automatically submitted at the end of time, which could mean the necessity for more time available on the quiz. Alternatively, it could also be a sign of procrastination, as students might have finished the quiz but expect it to be automatically submitted. Therefore, it would be necessary and beneficial to verify whether the students managed to answer all the questions or needed more time.

Still, in the same model, it was perceived that some students consulted the resources page immediately before or during the execution of the quizzes. This information can be convenient in courses where the exam must be taken without consultation, and therefore viewing other pages during the exam is a violation of the exam rules, which could be prevented by implementing tools that lock down the testing environment within a learning management system. Beyond that, it displays a multitasking behavior considering that the student is in another activity or page while performing the quiz.

From the comparison of the process map of all students with the map of passing students, it could be observed how similar they are. This highlights the inefficiency of the first model in representing unusual behaviors and, in particular, in noticing students with more difficulties or less motivated, that is, students at risk of failing.

Concerning the processes that were originated from the clusters, we know that Cluster 0 has two more students than Cluster 1; however, it is visible through the processes and the overall statistics that students in Cluster 1 are more active and interact more frequently with Moodle. Nevertheless, Cluster 0 students do not demonstrate a behavior similar to failing students or students at risk of failing. On the contrary, the lesser interaction in Moodle can mean that they are more direct in their search. Reflecting on the connections of the two cluster processes, we noticed the paths in Cluster 0 are shorter and link activities closer than those in Cluster 1. Therefore, it appears the students in Cluster 0 straightforwardly access the page they are searching in Moodle while Cluster 1 students navigate more from page to page, corroborating what was previously suggested that Cluster 0 students are more direct on their Moodle exploration.

In relation to Cluster 2, we managed to have represented a significant part of the students who failed. Nonetheless, it corresponds to the students who least interacted with Moodle, due to how clustering was executed, in other words, students that fail due to lack of work or interest. Therefore, the cluster-based process models concerning students at risk of failing only exemplify the behavior of students who demonstrate lower presence and frequency in Moodle. Furthermore, from the process model of Cluster 2, a distinction of the quiz activities can be identified. This can be a sign of a more significant presence on quiz days and a lack of commitment throughout the semester, as well as less frequent visits of other Moodle pages, particularly the consultation of available materials and participation in the course's forum.

After analyzing all these different process models, with varying levels of abstraction, it was even more notorious that the process model from the data set of all students did not represent the diverse behavior undertaken by the different types of students.

After a detailed analysis of the process models for the 2019/2020 school year, the application of the same strategy in the event logs of the students of the 2020/2021 academic year followed, particularly the application of the clustering technique and the process discover through Disco to test and evaluate its ability to predict the students' success. This made it possible to compare students undertaking the same course but in different academic years. However, in this case, we are not aware of the number of students who failed, nor is it possible to identify them. Hence, in the process models of all students from the two years (Figures 11 and 19), it was noted that in the year 2020/2021 the forum feature was less used in comparison with the previous year, which may reveal less collaboration between students or less interaction with the teacher.

Concerning the clusters of the data set of the academic year 2020/2021, we have two groups that contrast widely, Cluster 1 and 3. The former stands out for its significant interaction and activity in Moodle and the latter for its little presence and participation when browsing the course. This contrast is well represented in the process maps of each Cluster and highlights the insufficient work and effort of students belonging to Cluster 3 throughout the semester.

On the other hand, the process models of Cluster 2 and 3 present similarities despite the difference in the number of students, revealing the proximity of behavior upon visiting Moodle and possibly in the final results of these students. Therefore, in addition to having to pay attention to the trajectory of the students in Cluster 3, it seems that the students of Cluster 2 may also need special attention or an incentive due to the similar behavior that conveys lack of motivation or learning difficulties.

Furthermore, by contemplating and comparing the process models of Clusters 2 and 3 of the academic year 2020/2021, as well as the processes of the failing student and Cluster 2 from 2019/2020, it is possible to notice evident resemblances in the processes. This allows us to conclude that there is a common pattern in students who are likely to fail or have a more challenging time succeeding in the course. Therefore, it is possible and advisable to analyze and compare the process models that represent students' activity in Moodle to predict students at risk of failing or low performing students and thus act in advance and provide feedback.

#### **4.6.2. Improvement Proposals**

Deriving from the analysis and previous discussion, it follows some recommendations to enhance the process of accessing and searching in Moodle and improve the learning experience for students and teachers.

As a first step, it is critical to simplify the process in order to facilitate the analysis and understanding of students' behavior as well as help students not diverge when consulting Moodle and thus focus on what is essential. Therefore, it is recommended to exclude irrelevant activities, such as those excluded in the analysis carried out earlier; this would make accessing and searching for materials more straightforward. Moreover, the elimination of the entries recorded in Moodle that do not lead to further activities, that is, when the students only enter the main page of the course in Moodle and then leave.

On the other hand, some activities proved to be significant enough to be considered mandatory, such as the *quiz summary viewed* activity, which, as we saw in Figure 15, was performed by 73.6% of the students who passed but only 27.3% by those who failed. As mentioned earlier, the absence of this activity may represent the lack of time to complete the quiz. However, it may also be the case that students skip the activity by going directly to the quiz submission without considering reviewing the quiz. Thus, the suggestion to make the *quiz summary viewed* as a mandatory activity allows students to be more thoughtful and review their answers before submitting the quiz.

In addition, the forum activities are contemplated and understood as essential activities despite having a minor frequency of students visits according to our models. Therefore, to reverse this trend, we advise the creation of forum discussions about class topics and assign extra points for students who participate in the online discussion, which would also serve as an incentive for students to visit the Moodle page more and access available materials.

As we have seen throughout the analysis, there are two types of students who fail: the ones represented through the clustering algorithm, such as those belonging to Cluster 2 of the 2019/2020 academic year, who show infrequently activity and interaction in Moodle; and those who failed despite having similar behavior to the majority that passed, which integrate Clusters 0 and 1. Therefore, to reduce the number of students who fail, we advise implementing training quizzes in Moodle. In these quizzes, students can practice and prepare for the exam, but the result does not count towards the final grade. In this way, students who failed despite being active in Moodle can test their knowledge, understand their difficulties and better prepare for the exam. Additionally, these quizzes could also be applied to encourage the less active students to consult the materials over the weeks and not just closer to the exam date. This could be achieved by making it mandatory to complete small training quizzes about previous subjects every week to access the subject and materials given in the current week.

## 5. CONCLUSION

### 5.1. SYNTHESIS OF THE RESEARCH

The application of process mining in the education domain is an emergent subject due to the potential to analyze and discover educational process models that can reproduce the students' behavior and the learning environment and, consequently, obtain information and facilitate the understanding of the educational process. Therefore, this dissertation started with an investigation into two major subjects, process mining and education. This was followed by extensive research on the merger of the two topics in the systematic literature review, which was conducted by studying existing resources related to process mining in education, which confirmed the critical and promising role of process mining in its application in education.

This study illustrated the benefits of combining process mining with clustering by integrating other types of data from the students' interaction in Moodle. The convenience of this, apart from generating models that are easier to interpret, is that it results in more detailed and accurate models and, therefore, enhanced process discovery performance and the comprehension of the resulting models. In addition, it allows the detection of specific behaviors such as consulting materials during the quiz, skipping activities, multitasking behavior and the navigation paths performed by the students.

With the detailed analysis of the process models, we were able to confirm the link between student behavior in Moodle and course success, as well as the possibility to detect students at risk of failing. Nonetheless, it is necessary to highlight the difficulty in detecting students who failed with a high presence and interaction in Moodle, since through the clustering algorithm, it was only possible to gather the failed students who presented less activity in the Moodle search.

Throughout this study, we acquired enough knowledge and extracted valuable information from the process models to answer the three research questions. RQ1: How effective is process mining in identifying specific student behavior patterns in the Learning Management System Moodle? Through the application of process mining, we obtained process models that provide a general idea of the students' actions; however, they are incapable of specifying or representing the students' diverse behavior while consulting Moodle. Nevertheless, this weakness was overcome with the use of clustering techniques. RQ2: How does process mining contribute to predicting which students will have the most difficulty in succeeding? The question was answered in section 4.5. Evaluation, where using the same strategy previously applied, both in the application of the clustering algorithm and in the subsequent process discovery through Disco, it was possible to distinguish students with identical modeling behavior. The comparison of the process models with those previously observed shows the possibility of distinguishing the same behavior in the Moodle search of students at risk of failing through resemblances in the process. From a practical point of view, using the information of which students could fail or have difficulties in the content of the course enables the instructors to provide feedback and act in advance. RQ3: From the process models analysis, is it possible to withdraw valuable recommendations to students and teachers towards a more personalized educational process? Not only with the analysis of the various process models, but also with the entire development of this study, from the necessary research for the literature review to the data exploration and the discussion of the results. All the different phases undertaken in this study were necessary to consolidate the

knowledge and realize some changes that could be beneficial for the teaching activity, as suggested at the end of the discussion section.

## **5.2. LIMITATIONS AND FURTHER RESEARCH**

The limitation of this dissertation, which should direct and serve as a starting point for future research, can be distinguished based on the nature of the data or the methodological choices that may have influenced the results. We ought not assume that all possibilities have been considered in any data mining or process mining context. Beyond that, it is unrealistic to suppose that every option is represented in the dataset and the resulted models.

Even though we have tried to maintain a decent level of granularity in the process discovery procedure, future research should consider other levels of granularity in order to incorporate different user behavior patterns. Furthermore, it is necessary to acknowledge that the event logs contain only sample behavior. If something was not recorded in the event log and is therefore not represented in the process models, it does not mean that it did not happen in reality. This is even more significant in small event logs with a lot of variability, which is the case under study.

Additional limitations worth noting are the lack of information about the course. It would be beneficial to know further details in future works, such as the type of evaluation undertaken, students' characteristics, performance and final grades in contemplation of a more enriching analysis. In the same perspective, regarding the variables chosen for the clustering algorithm, the instructors of each course would comprehend and select more appropriate and crucial variables in order to represent successful performance in their course than the ones employed.

The present study concentrates on the analysis of only one course; further research can focus on other courses or groups of courses, aggregated by degree, academic year or area, such as management sciences, informatics or mathematics. There may also be an investigation focusing on other services available at the university or in Moodle. In addition, there is the possibility of going further with the study through the use of the other two techniques of process mining, conformance checking and process enhancement. Not least, it is necessary to consider and address other relations besides the current one that correlates success to the behavior in Moodle.

Finally, it is necessary to mention that this dissertation was elaborated in the context of the COVID-19 pandemic. Therefore, while we do not expect significant changes in students' behavior when interacting with Moodle, except for the participation in online exams, it would also be interesting to conduct further research in a post-pandemic context.

## REFERENCES

- Aguirre, S., Parra, C., & Sepúlveda, M. (2017). Methodological proposal for process mining projects. *International Journal of Business Process Integration and Management*, 8(2), 102-113. <https://doi.org/10.1504/IJBPIIM.2017.083793>
- Álvarez, P., Fabra, J., Hernández, S., & Ezpeleta, J. (2016). *Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs*. Paper presented at the 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET). <https://doi.org/10.1109/ITHET.2016.7760720>
- Aulia, D., & Waspada, I. (2019). The design of exploratory application and preprocessing of event log data in lms moodle-based online learning activities for process mining. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, 5(2), 124-133. <https://doi.org/10.23917/khif.v5i2.8023>
- Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *WIREs Data Mining and Knowledge Discovery*, 8(1), e1230. <https://doi.org/10.1002/widm.1230>
- Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). *Clustering for improving educational process mining*. Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, Indianapolis, Indiana, USA. <https://doi.org/10.1145/2567574.2567604>
- Bolt, A., de Leoni, M., & van der Aalst, W. M. P. (2016). Scientific workflows for process mining: building blocks, scenarios, and implementation. *International Journal on Software Tools for Technology Transfer*, 18(6), 607-628. <https://doi.org/10.1007/s10009-015-0399-5>
- Bolt, A., de Leoni, M., van der Aalst, W. M. P., & Gorissen, P. (2017). *Business Process Reporting Using Process Mining, Analytic Workflows and Process Cubes: A Case Study in Education*. Paper presented at the International Symposium on Data-Driven Process Discovery and Analysis, Springer, Cham. [https://doi.org/10.1007/978-3-319-53435-0\\_2](https://doi.org/10.1007/978-3-319-53435-0_2)
- Bozkaya, M., Gabriels, J., & Werf, J. M. v. d. (2009). *Process Diagnostics: A Method Based on Process Mining*. Paper presented at the 2009 International Conference on Information, Process, and Knowledge Management. <https://doi.org/10.1109/eKNOW.2009.29>
- Bruner, J. S. (1977). *The Process of Education: A landmark in educational theory*. Harvard University Press.
- Cairns, A. H., Gueni, B., Assu, J., Joubert, C., & Khelifa, N. (2015). *Analyzing and improving educational process models using process mining techniques*. Paper presented at the IMMM 2015 Fifth International Conference on Advances in Information Mining Management.
- Cairns, A. H., Gueni, B., Fhima, M., Cairns, A., David, S., & Khelifa, N. (2015). *Process mining in the education domain*. Paper presented at the International Journal on Advances in Intelligent Systems.
- Celonis (2020). Process Mining and Execution Management Software | Celonis. Retrieved from <https://www.celonis.com/>

- Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, 32(1), 74-88.  
<https://doi.org/10.1007/s12528-019-09225-y>
- Corallo, A., Lazoi, M., & Striani, F. (2020). Process mining and industrial applications: A systematic literature review. *Knowledge and Process Management*, 27(3), 225-233.  
<https://doi.org/10.1002/kpm.1630>
- Dacic, D., Stefanovic, D., Cosic, I., Lolic, T., & Medojevic, M. (2018). Business Process Mining Application: A Literature Review. In *Proceedings of the 29th International DAAAM Symposium 2018* (Vol. 29).
- Dictionary.com. (n.d.). Education. In *Dictionary.com* dictionary. Retrieved March 14, 2021, from <https://www.dictionary.com/browse/education>
- Dolak, R. (2019). *Using Process Mining Techniques to Discover Student's Activities, Navigation Paths, and Behavior in LMS Moodle*. Paper presented at the International Conference on Innovative Technologies and Learning. Springer, Cham. [https://doi.org/10.1007/978-3-030-35343-8\\_14](https://doi.org/10.1007/978-3-030-35343-8_14)
- dos Santos Garcia, C., Meincheim, A., Junior, E. R. F., Dallagassa, M. R., Sato, D. M. V., Carvalho, D. R., ... Scalabrin, E. E. (2019). Process mining techniques and applications – A systematic mapping study. *Expert Systems with Applications*, 133, 260-295.  
<https://doi.org/10.1016/j.eswa.2019.05.003>
- Dunzer, S., Stierle, M., Matzner, M., & Baier, S. (2019). *Conformance checking: a state-of-the-art literature review*. Paper presented at the Proceedings of the 11th International Conference on Subject-Oriented Business Process Management. <https://doi.org/10.1145/3329007.3329014>
- ECSSR (2019). *Education and Development: Investing in the Future*. The Emirates Center for Strategic Studies and Research. [https://www.ecssr.ae/en/publication/education\\_investing\\_in\\_future/](https://www.ecssr.ae/en/publication/education_investing_in_future/)
- Fluxicon (2020). Process Mining in Practice – Process Mining Book 2.5. Retrieved from <https://fluxicon.com/book/read/>
- Fluxicon (2021). Process Mining and Automated Process Discovery Software for Professionals – Fluxicon Disco. Retrieved from <https://fluxicon.com/disco/>
- Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15), 2479-2481. <https://doi.org/10.1093/bioinformatics/bth261>
- Graafmans, T., Turetken, O., Poppelaars, H., & Fahland, D. (2021). Process Mining for Six Sigma. *Business & Information Systems Engineering*, 63(3), 277-300. <https://doi.org/10.1007/s12599-020-00649-w>
- Günther, C. W., & van der Aalst, W. M. P. (2007). *Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics*. Paper presented at the International Conference on Business Process Management, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-75183-0\\_24](https://doi.org/10.1007/978-3-540-75183-0_24)

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.*, 11(1), 10–18.  
<https://doi.org/10.1145/1656274.1656278>
- Intayoad, W., Kamyod, C., & Temdee, P. (2018). *Process mining application for discovering student learning paths*. Paper presented at the 2018 International Conference on Digital Arts, Media and Technology (ICDAMT). <https://doi.org/10.1109/ICDAMT.2018.8376527>
- Jornitz, S., & Engel, L. C. (2021). The Management and Use of Data in Education and Education Policy: Introductory Remarks. *International Perspectives on School Settings, Education Policy and Digital Strategies: A Transatlantic Discourse in Education Research*, 223.
- 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>
- Kumar, S., & Ahmad, S. (2008). Meaning, aims and process of education. *School of Open Learning*, 3-6.
- Leno, V., Armas-Cervantes, A., Dumas, M., Rosa, M. L., & Maggi, F. M. (2018). *Discovering process maps from event streams*. Paper presented at the Proceedings of the 2018 International Conference on Software and System Process. <https://doi.org/10.1145/3202710.3203154>
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., ... Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, 62(10), e1-e34. <https://doi.org/10.1016/j.jclinepi.2009.06.006>
- Ly, L. T., Maggi, F. M., Montali, M., Rinderle-Ma, S., & van der Aalst, W. (2013). *A Framework for the Systematic Comparison and Evaluation of Compliance Monitoring Approaches*. Paper presented at the 2013 17th IEEE International Enterprise Distributed Object Computing Conference. <https://doi.org/10.1109/EDOC.2013.11>
- Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior*, 80, 179-196.  
<https://doi.org/10.1016/j.chb.2017.11.011>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Mukala, P., Buijs, J., Leemans, M., & van der Aalst, W. (2015). *Learning Analytics on Coursera Event Data: A Process Mining Approach*. Paper presented at the Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2015).
- Munoz-Gama, J., & Carmona, J. (2011). *Enhancing precision in Process Conformance: Stability, confidence and severity*. Paper presented at the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). <https://doi.org/10.1109/CIDM.2011.5949451>

- Nafasa, P., Waspada, I., Bahtiar, N., & Wibowo, A. (2019). *Implementation of Alpha Miner Algorithm in Process Mining Application Development for Online Learning Activities Based on MOODLE Event Log Data*. Paper presented at the 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). <https://doi.org/10.1109/ICICoS48119.2019.8982384>
- OECD (2010). Context and main concepts. In *Recognising Non-Formal and Informal Learning: Outcomes, Policies and Practices*. Paris: OECD Publishing.
- OECD (2019). *Trends Shaping Education 2019*. Paris: OECD Publishing.
- ProM Tools (n.d.). Retrieved from ProM Tools: <http://www.promtools.org/doku.php>
- RapidMiner (2021). Expectation Maximization Clustering – RapidMiner Documentation. Retrieved from [https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/expectation\\_maximization\\_clustering.html](https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/expectation_maximization_clustering.html)
- RapidProM (n.d.). Retrieved from RapidProM: <http://www.rapidprom.org/>
- Real, E. M., Pimentel, E. P., de Oliveira, L. V., Braga, J. C., & Stiubiener, I. (2020). *Educational Process Mining for Verifying Student Learning Paths in an Introductory Programming Course*. Paper presented at the 2020 IEEE Frontiers in Education Conference (FIE). <https://doi.org/10.1109/FIE44824.2020.9274125>
- Reinkemeyer, L. (2020). Process Mining in a Nutshell. In L. Reinkemeyer (Ed.), *Process Mining in Action: Principles, Use Cases and Outlook* (pp. 3-10). Springer, Cham. [https://doi.org/10.1007/978-3-030-40172-6\\_1](https://doi.org/10.1007/978-3-030-40172-6_1)
- Rix, J., & Twining, P. (2007). Exploring education systems: towards a typology for future learning? *Educational Research*, 49(4), 329-341. <https://doi.org/10.1080/00131880701717180>
- Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61, 224-236. <https://doi.org/10.1016/j.jbi.2016.04.007>
- Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: a tutorial and case study using Moodle data sets. In *Data Mining and Learning Analytics*, 1-28.
- Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1), 12-27. <https://doi.org/10.1002/widm.1075>
- Rovani, M., Maggi, F. M., de Leoni, M., & van der Aalst, W. M. P. (2015). Declarative process mining in healthcare. *Expert Systems with Applications*, 42(23), 9236-9251. <https://doi.org/10.1016/j.eswa.2015.07.040>
- Saint, J., Gašević, D., & Pardo, A. (2018). *Detecting Learning Strategies Through Process Mining*. Paper presented at the European Conference on Technology Enhanced Learning, Springer, Cham. [https://doi.org/10.1007/978-3-319-98572-5\\_29](https://doi.org/10.1007/978-3-319-98572-5_29)

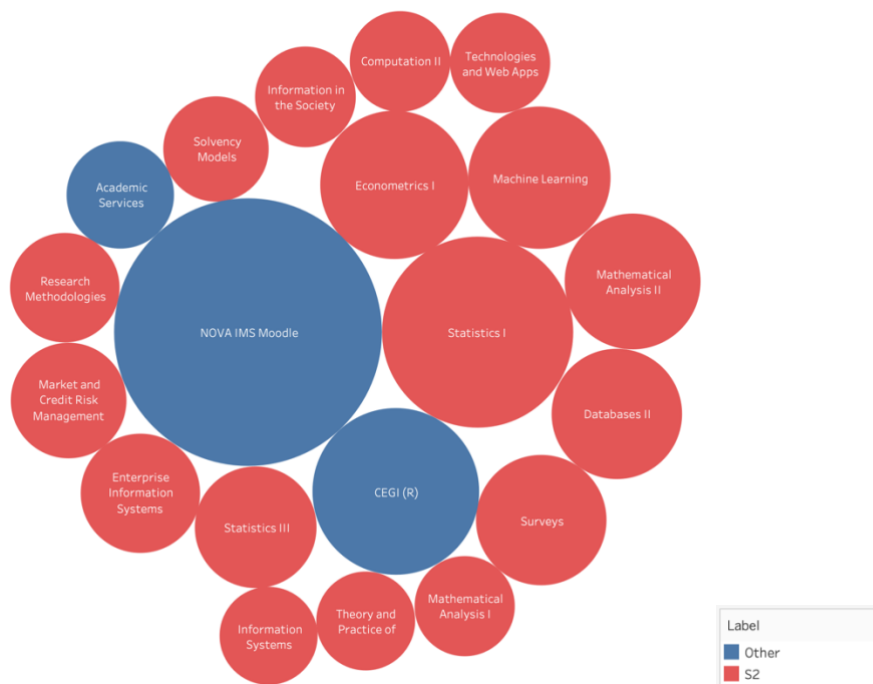
- Saint, J., Whitelock-Wainwright, A., Gašević, D., & Pardo, A. (2020). Trace-SRL: A Framework for Analysis of Microlevel Processes of Self-Regulated Learning From Trace Data. *IEEE Transactions on Learning Technologies*, 13(4), 861-877. <https://doi.org/10.1109/TLT.2020.3027496>
- Salazar-Fernandez, J. P., Sepúlveda, M., & Munoz-Gama, J. (2019). *Influence of Student Diversity on Educational Trajectories in Engineering High-Failure Rate Courses that Lead to Late Dropout*. Paper presented at the 2019 IEEE Global Engineering Education Conference (EDUCON). <https://doi.org/10.1109/EDUCON.2019.8725143>
- Sedrakyan, G., De Weerd, J., & Snoeck, M. (2016). Process-mining enabled feedback: “Tell me what I did wrong” vs. “tell me how to do it right”. *Computers in Human Behavior*, 57, 352-376. <https://doi.org/10.1016/j.chb.2015.12.040>
- Selwyn, N. (2015). Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1), 64-82. <https://doi.org/10.1080/17439884.2014.921628>
- Song, M., Günther, C. W., & van der Aalst, W. M. P. (2009). *Trace Clustering in Process Mining*. Paper presented at the International Conference on Business Process Management, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-00328-8\\_11](https://doi.org/10.1007/978-3-642-00328-8_11)
- Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills. In *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*: OECD Publishing, Paris. <https://doi.org/10.1787/9789264273955-14-en>
- Trcka, N., & Pechenizkiy, M. (2009). *From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining*. Paper presented at the 2009 Ninth International Conference on Intelligent Systems Design and Applications. <https://doi.org/10.1109/ISDA.2009.159>
- Turner, D. A. (2009). *Theory and Practice of Education* (C. I. P. Group Ed. 1st ed.).
- Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning*, 10(2), 160-176. <https://doi.org/10.1108/JRIT-09-2017-0022>
- UNESCO Institute for Statistics (UIS) (2012). *International Standard Classification of Education, ISCED 2011*.
- UNESCO Institute for Statistics (UIS) (2014). *ISCED fields of education and training 2013 (ISCED-F 2013): Manual to accompany the International Standard Classification of Education 2011*.
- UNESCO International Bureau of Education (IBE) (2012). *General Education Quality Analysis/Diagnosis Framework*: UNESCO International Bureau of Education.
- UNESCO International Bureau of Education (IBE) (2013). *Glossary of Curriculum Terminology*: UNESCO International Bureau of Education.
- Uzir, A., Gašević, D., Matcha, W., Jovanović, J., Pardo, A., Lim, L.-A., & Gentili, S. (2019). *Discovering time management strategies in learning processes using process mining techniques*. Paper

- presented at the European Conference on Technology Enhanced Learning.  
[https://doi.org/10.1007/978-3-030-29736-7\\_41](https://doi.org/10.1007/978-3-030-29736-7_41)
- van den Beemt, A., Buijs, J., & van der Aalst, W. (2018). Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering. *International Review of Research in Open and Distributed Learning*, 19(5).  
<https://doi.org/10.19173/irrodl.v19i5.3748>
- van der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes* (1st ed.): Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-19345-3>
- van der Aalst, W. (2016). *Process mining: Data science in action* (2nd ed.): Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-662-49851-4>
- van der Aalst, W. (2018). *Process mining and simulation: a match made in heaven!* Paper presented at the Proceedings of the 50th Computer Simulation Conference.  
<https://doi.org/10.22360/summersim.2018.scsc.005>
- van der Aalst, W. (2020). Academic View: Development of the Process Mining Discipline. In Reinkemeyer L. (Ed.), *Process Mining in Action: Principles, Use Cases and Outlook*, 181-196. Springer, Cham. [https://doi.org/10.1007/978-3-030-40172-6\\_21](https://doi.org/10.1007/978-3-030-40172-6_21)
- van der Aalst, W., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... Wynn, M. (2012). *Process Mining Manifesto*. Paper presented at the International Conference on Business Process Management, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-28108-2\\_19](https://doi.org/10.1007/978-3-642-28108-2_19)
- van der Aalst, W., Adriansyah, A., & van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. *WIREs Data Mining and Knowledge Discovery*, 2(2), 182-192. <https://doi.org/10.1002/widm.1045>
- van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2005). *The ProM Framework: A New Era in Process Mining Tool Support*. Paper presented at the International Conference on Application and Theory of Petri Nets, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/11494744\\_25](https://doi.org/10.1007/11494744_25)
- van Eck, M. L., Lu, X., Leemans, S. J. J., & van der Aalst, W. M. P. (2015). *PM<sup>2</sup>: A Process Mining Project Methodology*. Paper presented at the International Conference on Advanced Information Systems Engineering, Springer, Cham. [https://doi.org/10.1007/978-3-319-19069-3\\_19](https://doi.org/10.1007/978-3-319-19069-3_19)
- Vidal, J. C., Vázquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems*, 100, 160-174.  
<https://doi.org/10.1016/j.knosys.2016.03.003>
- Weijters, A., van Der Aalst, W. M., & De Medeiros, A. A. (2006). Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP*, 166, 1-34.

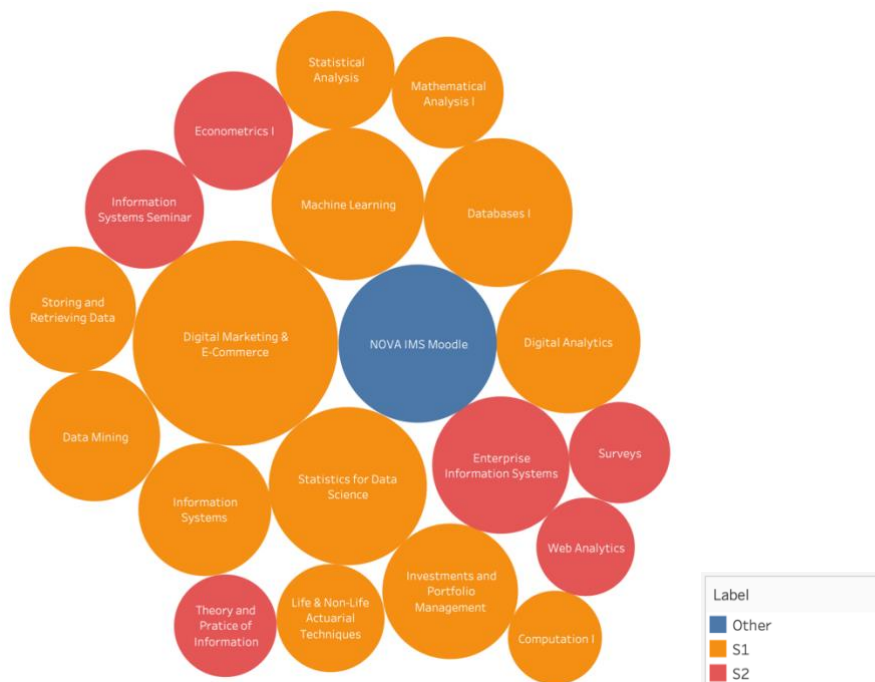
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). The WEKA Workbench. In *Data Mining: Practical Machine Learning Tools and Techniques* (Fourth Edition ed., 553-571): Morgan Kaufmann. [https://www.cs.waikato.ac.nz/ml/weka/Witten\\_et\\_al\\_2016\\_appendix.pdf](https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf)

## APPENDICES

### Appendix I

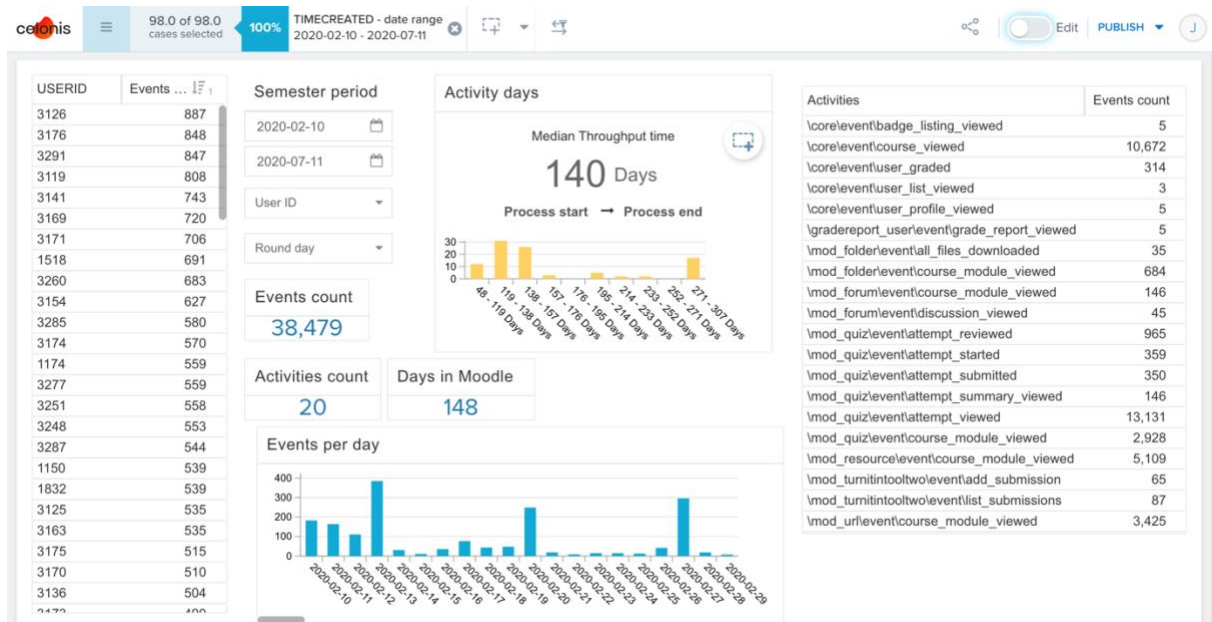


Packed bubble chart of the 20 courses with the most events in 2020

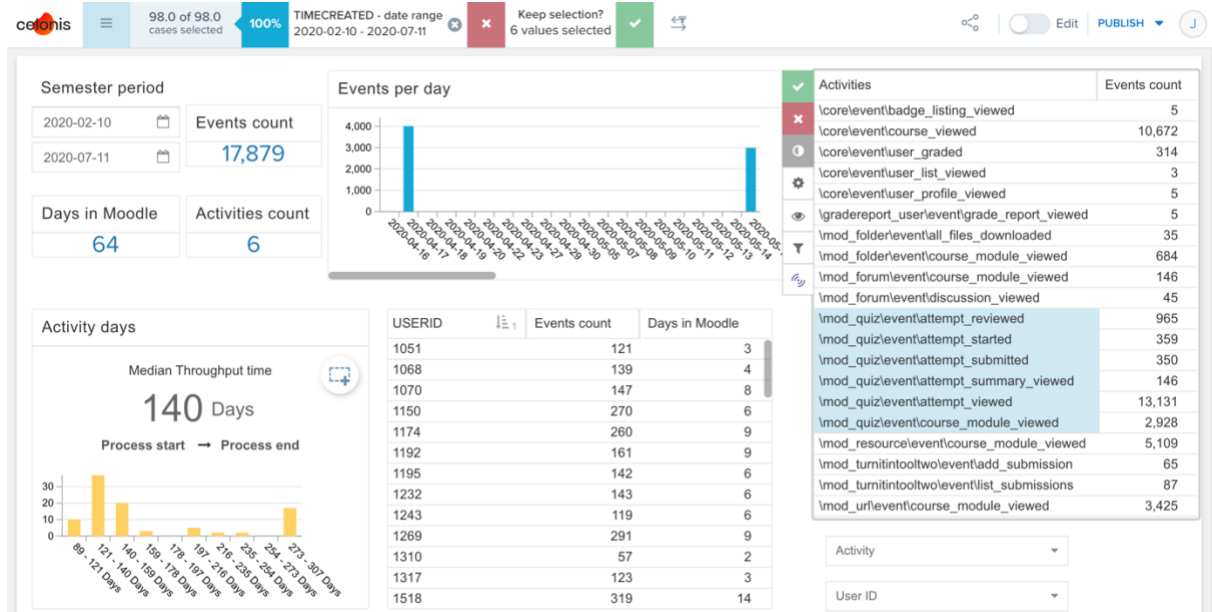


Packed bubble chart of the 20 courses with the most events in 2021

## Appendix II



Celonis overview board



Celonis overview board with selection of quiz activities

