Process parameters discovery based on application of k-means algorithm - a real case experimental study

Snježana Križanić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin
skrizanic@foi.hr

Katarina Tomičić-Pupek
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin
katarina.tomicic@foi.hr

Abstract. *This paper describes the application of the k-means algorithm on educational data to detect processes in the e-learning environment at a Higher Education institution in Croatia. Based on the interpretation of analysed data sets, two illustrative examples of process models are given, showing time reference and activity frequency as valuable process parameters extractable from raw data. The goal of this paper is to explore the application of a grouping method over educational logs with the aim to identify potential process parameters needed to understand student behaviour in accessing teaching material.*

Keywords. process discovery, process parameters, k-means, educational event logs

1 Introduction

Process mining consists of a set of techniques that automatically constructs a model of current organization activity based on event logs (Sarno, Effendi & Haryadita, 2016). The aim of the process discovery is to facilitate the discovery process models elements that precisely describe processes by considering only records of an organization which describe its operational processes. This kind of records are usually captured in the form of event logs, which consist of cases and events related to these cases, and are used to enrich process models (Buijs, Van Dongen & Van der Aalst, 2014; Van der Aalst, Adriansyah & Van Dongen, 2012). Using such event logs, we can detect the organization's processes that are performed in the organization. By modelling recognized processes, preconditions for advanced analyses (e.g., simulations) are created, possible process improvements can be detected, as well as potential issues in the existing model which need to be addressed by introducing improvements or innovations into the process.

The aim of this paper is to investigate student behaviour in the processes of using an e-learning system, based on a set of event logs and with application of the k-means algorithm for grouping. K-means algorithm is applied for identifying groups of students with similar behaviour in order to address following research questions:

1. Which variables can be extracted from event logs of an e-learning system?
2. Which variable values had a significant influence on grouping students by their behaviour in using e-learning systems?
3. How does grouping students by their behaviour impact process parameters?

Chapter 2 represents literature review about process mining and process discovery. Chapter 3 represents the methodology of the research. Chapter 4 describes the research results, followed by a short conclusion on our research.

2 Literature review

By using process mining, organizations can discover processes as they were conducted in reality, check if certain practices and regulations were really followed, and get "insight into bottlenecks, resource utilization, and other performance-related aspects of processes" (de Leoni, Van der Aalst & Dees, 2016). The traditional goal of process mining is to find a process models which describe the system as accurately as possible, using only the observed behaviour of the event logs (Buijs, Van Dongen & Van der Aalst, 2014).

For comparing and evaluating detected process model, different algorithms can be used (Sarno, Effendi & Haryadita, 2016). "Fitness, precision, generalization and structure are used to evaluate the quality of a model" (Sarno, Effendi & Haryadita, 2016). Buijs, van Dongen & van der Aalst (2014) state that four dimensions of quality: simplicity, convenience, precision and generalization can be considered when determining the quality of the process model, where simplicity is a model property, and three other dimensions refer to event records. García-Bañuelos et al. (2014) agree that numerous algorithms for automatic process detection have been developed

with the aim of balancing the compromise between precision, generalization and simplicity of the detected models.

Automatic discovery techniques can also be used to reconstruct the process models (Dumas et al., 2013, p. 162). Once the correct business process model is discovered, it is important to test its validity and to optimize the model fitness, whereby fitness quantifies how much of the observed behaviour is covered by the model (Sarno, Effendi & Haryadita, 2016; Bose & Van der Aalst, 2010). Fitness is measured as a value between 0 and 1, whereby the value 1 indicates perfect fitness, and the value 0 indicates very poor fitness. If the model fitness is closer to 1, it means that alignment contains only synchronous moves and no deviations. "As the percentage of moves on log and model over all moves increases, the fitness value decreases" (de Leoni, Van der Aalst & Dees, 2016). A process model with a good fitness "fits well with the reality; all the traces in the event logs are represented in the discovered process model" (Sarno, Effendi & Haryadita, 2016).

Literature shows that several measures can be defined, which point to the quality of the process model in a given dimension. However, there are some important requirements that every measure should follow: efficient implementation, intuitive results, clear specification and orthogonality (Buijs, Van Dongen & Van der Aalst, 2014). The same authors also indicate that process trees are often used as a convenient way of representing the process model, and de Leoni, & van der Aalst and Dees (2016) agree with that assumption.

In education, process mining can be used to identify activities ran by students while they are preparing for lectures and exams. Using process mining, it is possible to identify frequency and timing parameters of activities and interpret this data during process discovery.

Dutt, Ismail & Herawan (2017) state that "an educational institution environment broadly involves three types of actors namely teacher, student and the environment". By Campagni et al. (2015), it is important to explore and analyse "the information stored in student databases in order to understand and improve the performance of the student learning process". Data clustering used with k-means algorithm enables "academicians to predict student performance, associate learning styles of different learner types and their behaviour and collectively improve upon institutional performance" (Dutt, Ismail & Herawan, 2017).

There are many databases containing student's information, so we can operate with large repositories of data reflecting how students learn. E-learning environments continuously generate "large amounts of data concerning the interactions between teaching and learning" (Campagni et al., 2015). By extracting information from data, it is possible to generate process models representing various process scenarios in education.

3 Methodology

In this section, the data source, the data type, and the information contained in the data are described.

The data used for analysis are event logs from an e-learning system for one e-course at a Higher Education institution in Croatia for a student generation in 2018/2019. The time span within we observed data was from the February 2019 till June 2019. The logs were downloaded from the e-learning system and originally had 50.185 records for generation 2018/2019. The total number of students participating in the course was 213.

The raw data consisted of: access date and time, full name of the student, context (e.g. lecture, forum), the component (e.g. "record", "system"), description (e.g. "Log report viewed"), source (e.g. "web"), and the IP address of the student who accessed to e-course. Due to sensibility of data and privacy, only a subset of anonymized data was extracted for further analysis.

In relevance to our first research question data was analysed and from this subset of data following variables were recognized as significant to process parameters:

a) Access date and time for event logs corresponds to process parameters of activity occurrence in terms of frequency and timing;
b) Context from the event logs forms process scenarios consisting of different sequence of activities;
c) Description from the event logs supplements activity data with roles of process participants.

Process scenarios consisting of different sequence of activities are in our case the most desirable result. Therefore, variable named as "context" was selected for student grouping. Its values were: accessing lecture materials, accessing auditory exercises materials, accessing laboratory exercises materials, and visiting forums. The combination of these values, with the frequency based on access date and time, was significant for establishing process scenarios of student behaviour in using e-courses.

After forming a subset of data, pivot tables were created. These tables were imported in the RapidMiner tool ("RapidMiner Studio, Visual workflow designer for the entire analytics team," 2019) and a grouping analysis was performed. Pivot tables were designed to show the frequency of access of each student to materials from lectures, auditory exercises, laboratory exercises, and forums (news forum and discussion forum).

The tool settings were: the grouping variable was student's ID, the method chosen for normalization was Z-transformation, algorithm chosen for grouping was k-means, the number of groups was 3, because, according to testing with different number of clusters, it was considered as the best value with promising results. Due to the common use of clustering as data structure discovery tool, for a real application in improvement initiatives results with several clustering methods and parameters (e.g., selecting different

dissimilarities, different methods, different number of clusters if using k-means approach) should be explored. Measure types for grouping in our case were Numerical Measures and chosen numerical measure was Euclidean Distance. The variables that were selected as influential on the grouping of students, were frequencies of access of students to materials from lectures, auditory exercises (AE), laboratory exercises (LE), forums (news forum and discussion forum).

4 Analysis results

After preparing the data, the grouping analysis was performed. In this section, grouping (clustering) of students is presented, as well as the interpretation of this data in relation to process discovery. By analysing student's behaviour in using an e-learning system, real case process scenarios in the teaching material management process can be build.

4.1 Results of grouping based on event logs

K-means algorithm was applied on data gathered from a platform for e-learning, used as a support system. After performing the data analysis, for the purpose of grouping students, two models showing the results of grouping students were created.

There were 213 students in total enrolled in the course. Course had two mid-term exams within the regular program, which were not exclusive in terms of passing one in order to get access to the second. Dates of mid-term exams were applied as a time reference for splitting the period of the analysis in two sets. In total, two data tables were created representing data a) before the 1st mid-term exam, and b) after 1st and before 2nd mid-term exam. The frequency of access indicates student's motivation to use teaching materials for various forms of preparation, due to the fact that the exam includes two types of questions in the exam: theory oriented (mainly from lecture materials) and practical assignments (mainly from AE and LE materials).

Figures 1 and 2 are showing the results of grouping the students of generation 2018/2019 before the first mid-term exam, which was written in April 2019. In the group 0, there are 78 students, in the group 1 there are 115 students, and in the group 2, there are 20 students. According to the centroid table (Table 1), group 0 gathers students who had overall medium frequency of access to the content in the e-learning system, but mostly oriented on the practical assignments. They accessed to the auditory and laboratory exercises more often than to the materials from lectures and forums. Group 1 holds students with the lowest access to contents in the e-learning system. They accessed forums more often than other content. Group 2 includes the least number of students and, at the same time, they had the highest frequency of access to lecture materials, followed by access to forums, laboratory exercises, and auditory exercises.

Figure 3 shows the success rate via a box plot made by grades (scores) of the students. In the cluster (group) 0 are the grades of the first mid-term exam from the students which belong to group 0 by grouping analysis using the k-means algorithm. The cluster 1 contains grades of students which by Figure 1 belong to group 1. In the cluster 2 are the grades of the first mid-term exam from the students which by Figure 1 belong to the group 2. According to this representation, we can see that students from the group 1, which had the lowest frequencies of access to the materials from lectures, forums, auditory and laboratory exercises, also had the lowest grades from the first mid-term exam. In the group 2 are the students who accessed the content in the e-learning environment the most, and they also had the best results on the first mid-term exam.

Further on, it was interesting to explore whether the frequency of access to the specific content in the e-learning environment influenced the outcome of the first mid-term exam. Due to the types of assignment in the exam, both preparation strategies (theory or practical orientation) were used by students, whereby theory orientation was the most successful one because it contributed to the overall success rate in the exam.

Figures 4 and 5 are representing the results of grouping analysis till the second mid-term exam which was written in June 2019. There were 30 students in group 0, 163 students in group 1, and 14 students in group 2. According to the Table 2, in the group 0 are the students who accessed mostly materials from auditory exercises and then lectures. They haven't accessed the forums very often. Group 1 gathers students with the lowest frequency of access to the materials from lectures, forums, auditory and laboratory exercises. Group 2 contains students with the highest frequency of access to the forums. Students in the group 0 were more active in the overall access than the students from the group 2. Belonging to group 2, students showed the highest frequency to forums and the materials from laboratory exercises, but the students in the group 0 had higher frequency to the materials from lectures and auditory exercises.

Figure 6 represents the success rate via box plot according to the grades from the second mid-term exam of the students. Cluster 0 from the Figure 6 contains grades of students which belong to the group 0 by grouping analysis from the Figure 4. The cluster 1 from the Figure 6 contains the grades of the students which belong to group 1 from the Figure 4. Finally, the cluster 2 from the Figure 6 contains the grades of the students belonging to group 2 from the Figure 4. According to this representation, students from the group 0, made by grouping analysis, had the best grades at the second mid-term exam. Similar as the first mid-term exam results, again both preparation strategies (theory or practical orientation) contributed to the success rate in the exam.

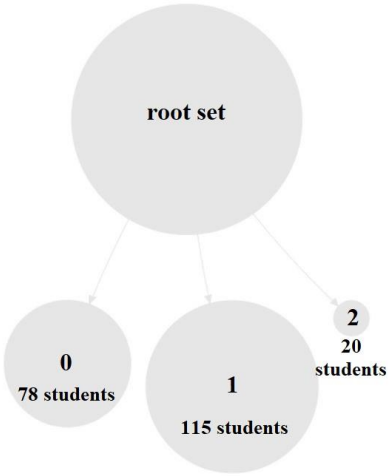

Figure 1 Tree, 1st mid-term exam, 2019

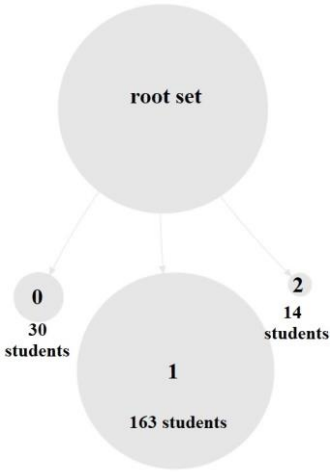

Figure 4 Tree, 2nd mid-term exam, 2019

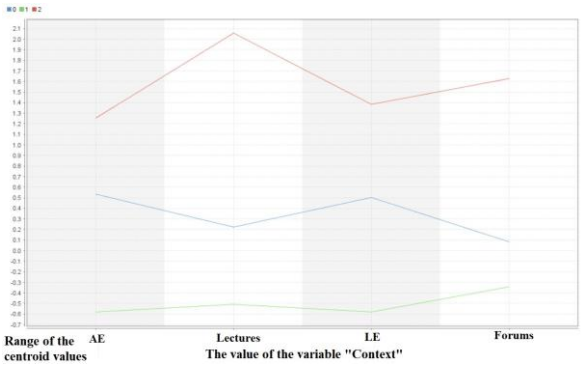

Figure 2 Plot, 1st mid-term exam, 2019

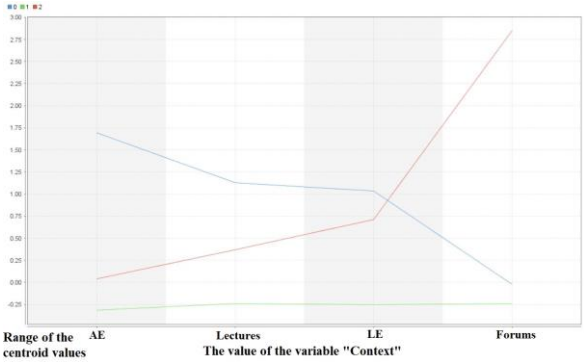

Figure 5 Plot, 2nd mid-term exam, 2019

Table 1. Centroid table for 1st mid-term exam

Attribute	Group_0	Group_1	Group_2
Auditory exercises	0.534	-0.580	1.254
Materials from lectures	0.224	-0.510	2.058
Laboratory exercises	0.502	-0.581	1.382
Forums	0.085	-0.341	1.629

Table 2. Centroid table for 2nd mid-term exam

Attribute	Group_0	Group_1	Group_2
Auditory exercises	1.691	-0.315	0.040
Materials from lectures	1.125	-0.239	0.367
Laboratory exercises	1.033	-0.251	0.711
Forums	-0.023	-0.241	2.849

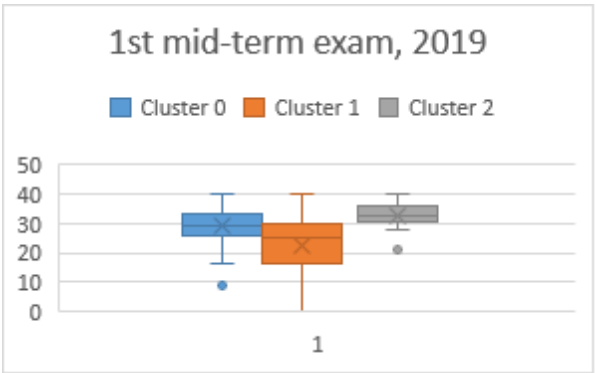

Figure 3 1st mid-term exam by grades, 2019

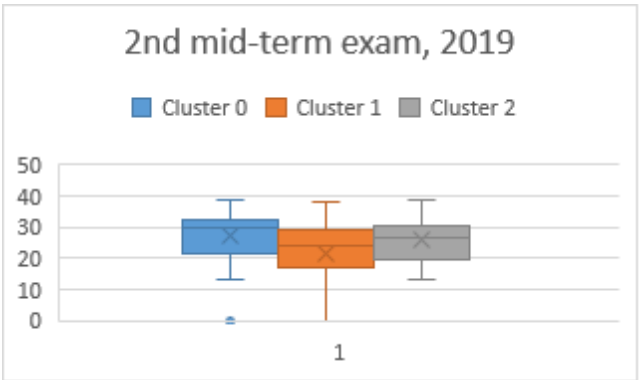

Figure 6 2nd mid-term exam by grades, 2019

4.2. Interpretation of data for process discovery

By interpretation of data about student's behaviour in using an e-learning system, relevant process parameters can be identified as a first step in process discovery based on data analysis from event logs. Two illustrations of the data interpretation are shown in models depicted in Figures 7 and 8, whereby the models have been created using Bizagi Modeler ("Bizagi - Digital Process Automation and BPM", 2019).

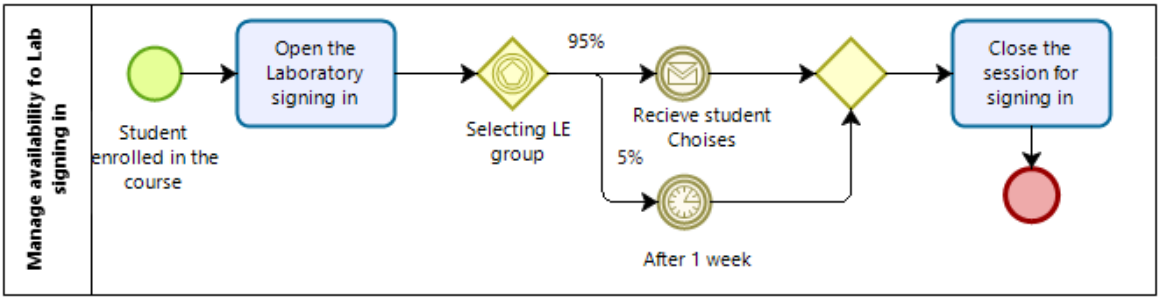

Figure 7 Managing availability to Lab signing in

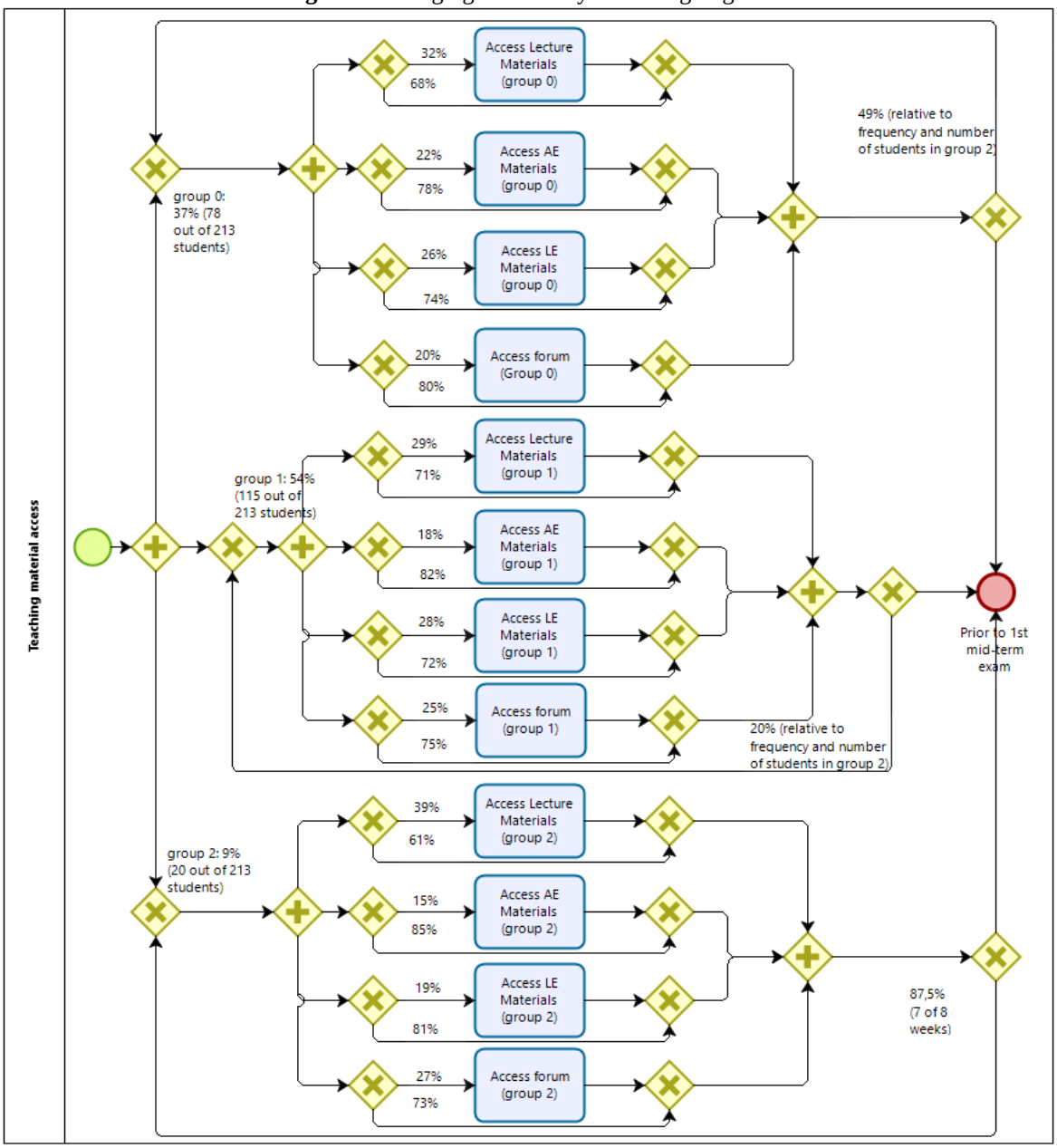

Figure 8 Teaching material access prior to the 1st mid-term exam

Figure 7 is giving an example of translating the time reference into the process parameter "Timer" for restricting availability for signing into a group for Laboratory exercises. In the beginning of the semester, all the students accessed the e-course searching for information about assessment model. Among students included in the regular study who had one week to decide which LE group they plan to attend, 95% of them have decided promptly, and others decided through the week.

Figure 8 shows the interpretation of material access frequency across groups of students into ratios of three scenarios in the process of using teaching materials in a real process. Learning materials were available to students mostly few days earlier, so that the students could use these materials before lectures in the current week. After the lectures, students should participate in auditory exercises, whereby students should have downloaded the materials from the e-learning system before the classes. After the AE, students participated in laboratory exercises (LE). During the LE students had to download the LE materials from the e-learning system.

The data about accessing teaching materials, illustrated in this analysis, allows an identification of probabilities for triggering activities in various sequences. These sequences of activities, and the rules for activity triggering, shape process scenarios that occur with various probability. Also, the time reference in the log data, can be interpreted as a time trigger for the process scenario which is another important process parameter.

By identifying potential process parameters for process scenarios analysis, more detailed understanding of how student's behaviour in accessing teaching material is influencing exam results, can be achieved. Based on this kind of analysis, improvement options can be proposed, e.g. cutting availability of teaching materials, merging them into single lecture preparation documents, giving credits for accessing and other options.

While discovering the most informative and the most separable clusters, we recognized the value of data for group 2, which shows that the behaviour of continuously accessing diverse course materials has contributed to the success rate in the 1st mid-term exam. In this experimental study, comparison of these groups in both stages shows that the common property is student's affinity for continuously accessing course materials in 49% and 87,5%, respectively. More detailed property analysis would require the application of more clustering methods and parameters on broader data sets.

5 Conclusion

In this paper, variables extracted from event logs of an e-learning system were analysed in order to explore which variable values had influence on grouping students by their behaviour in using e-learning systems. Based on the data acquired and by applying k-means algorithm for grouping students by their behaviour, significant process parameters can be identified. Interpretation of these process parameters is valuable for understanding process sequences and investigating its feasible scenarios.

According to the frequencies of access, in our real case experimental study, we analysed and created two illustrations of process models interpreting student's behaviour prior the first mid-term exam. Future work will be oriented on performing deeper analysis of event logs in order to generate a framework for evaluating process scenarios and thereby facilitating the proposition of improvements or innovations in the teaching material management process.

References

Bizagi - Digital Process Automation and BPM (2019). Retrieved from https://www.bizagi.com/.

Bose, R.P.J.C., & Van der Aalst, W.M.P. (2010). Trace clustering based on conserved patterns: Towards achieving better process models. *Lecture Notes in Business Information Processing*, vol. 43, 170-181.

Buijs, J. C. A. M., Van Dongen, B. F., & Van der Aalst, W. M. P. (2014). Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. *International Journal of Cooperative Information Systems*, 23(1).

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M.C. (2015). Data mining models for student careers. *Expert Systems with Applications*, 42(13), 5508-5521.

De Leoni, M., Van der Aalst, W. M. P., & Dees, M. (2016). A General Process Mining Framework for Correlating, Predicting and Clustering Dynamic Behavior Based on Event Logs. *Information Systems*, 56, 235-257.

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). *Fundamentals of Business Process Management*. Springer-Verlag, Berlin, Heidelberg.

Dutt, A., Ismail, M.A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, 5, 15991-16005.

García-Bañuelos, L., Dumas, M., La Rosa, M., De Weerdt, J., & Ekanayake, C.C. (2014). Controlled automated discovery of collections of business process models. *Information Systems*, 46, 85-101.

RapidMiner Studio, Visual workflow designer for the entire analytics team (2019). Retrieved from https://rapidminer.com/products/studio.

Sarno, R., Effendi Y. A., & Haryadita, F. (2016). Modified Time-Based Heuristic Miner for Parallel Business Processes. *International Review on Computers and Software*, 11(3), 248-260.

Van der Aalst, W., Adriansyah, A., & Van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2), 182-192.